AN INVITATION TO THE BOOTSTRAP: 
Panacea for Statistical Inference?

by Léopold SIMAR∗

HANDOUT

September 01, 2008

∗Comments from Ingrid Van Keilegom on a previous version of this handout are gratefully acknowledged.


## Contents

1 Basic Ideas

1.1 The Aim of Statistical Inference .......................... 1

1.2 The Bootstrap Principle .................................. 5

1.2.1 The Plug-in Principle ................................. 5

1.2.2 Computational aspects: Monte-Carlo Methods ....... 10

1.3 An Alternative to the Bootstrap: Asymptotic Theory .... 23

1.3.1 The Parametric Delta Method ......................... 23

1.3.2 Nonparametric Delta Method, Influence Functions ... 25

1.4 Bootstrap Estimation of Bias and Standard Errors ....... 29

1.4.1 Plug-in principle .................................... 29

1.4.2 Correction for the bias ............................... 32

1.5 Bootstrap Confidence Intervals .......................... 37

1.5.1 Basic bootstrap confidence intervals ................ 37

1.5.2 Practical computation of the quantiles $u^*(a)$ ....... 39

1.6 Summary Table of the Bootstrap Ideas .................. 45

2 More on Confidence Intervals .............................. 46

2.1 The Percentile Method .................................. 46

2.2 The BC (bias-corrected) Percentile Method ............... 49

2.3 The $BC_a$ Percentile Method ............................ 52

2.4 The Bootstrap-$t$ Method ................................ 56

2.4.1 Studentization in the Real World .................... 56

2.4.2 Studentization in the Bootstrap World ............... 57

2.5 Some Simulated Examples ............................... 60

2.6 Some Programs in MATLAB .............................. 71

2.6.1 Basic bootstrap functions ........................... 71

2.6.2 Example for $X$ in $Expo(\mu)$ ....................... 72

2.6.3 Monte-Carlo performance evaluation ................ 73
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 The Properties of the Bootstrap</td>
<td>76</td>
</tr>
<tr>
<td>3.1 Why the Bootstrap works</td>
<td>76</td>
</tr>
<tr>
<td>3.2 The Virtue of prepivoting</td>
<td>80</td>
</tr>
<tr>
<td>3.3 Consistency of the Bootstrap</td>
<td>82</td>
</tr>
<tr>
<td>3.4 Some Useful Tools for the Asymptotics</td>
<td>86</td>
</tr>
<tr>
<td>4 Hypothesis Testing</td>
<td>89</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>89</td>
</tr>
<tr>
<td>4.2 Parametric Bootstrap tests</td>
<td>92</td>
</tr>
<tr>
<td>4.3 Nonparametric Permutation Tests</td>
<td>96</td>
</tr>
<tr>
<td>4.4 Nonparametric Bootstrap Tests</td>
<td>108</td>
</tr>
<tr>
<td>5 The Bootstrap in Regression Models</td>
<td>116</td>
</tr>
<tr>
<td>5.1 Regression Models</td>
<td>116</td>
</tr>
<tr>
<td>5.2 Bootstrapping the Residuals</td>
<td>119</td>
</tr>
<tr>
<td>5.3 Monte-Carlo Performances of Confidence Intervals</td>
<td>121</td>
</tr>
<tr>
<td>5.4 Bootstrapping the Pairs</td>
<td>123</td>
</tr>
<tr>
<td>5.5 The Wild Bootstrap</td>
<td>128</td>
</tr>
<tr>
<td>5.6 Testing Hypothesis</td>
<td>131</td>
</tr>
<tr>
<td>5.7 Bootstrapped Prediction Intervals</td>
<td>135</td>
</tr>
<tr>
<td>6 Iterated Bootstrap</td>
<td>139</td>
</tr>
<tr>
<td>6.1 Estimating the variance for prepivoting</td>
<td>139</td>
</tr>
<tr>
<td>6.2 Calibrating coverage errors of CI</td>
<td>145</td>
</tr>
<tr>
<td>7 The Jacknife</td>
<td>153</td>
</tr>
<tr>
<td>7.1 The Jacknife method</td>
<td>153</td>
</tr>
<tr>
<td>7.1.1 Some examples</td>
<td>155</td>
</tr>
<tr>
<td>7.2 The Jacknife and the Bootstrap</td>
<td>158</td>
</tr>
<tr>
<td>7.3 The Jacknife and the Delta method</td>
<td>163</td>
</tr>
<tr>
<td>8 The Smoothed Bootstrap</td>
<td>165</td>
</tr>
<tr>
<td>8.1 Smooth Estimates of $F$</td>
<td>165</td>
</tr>
<tr>
<td>8.2 The bootstrap algorithm</td>
<td>169</td>
</tr>
<tr>
<td>8.3 Testing Unimodality</td>
<td>171</td>
</tr>
</tbody>
</table>
Chapter 1

Basic Ideas

The aim of this handout is to introduce to the basic ideas of the bootstrap and to show that it can be useful in several problems of Statistical Inference: bias and variance estimation, confidence intervals and testing hypothesis. It shows also how to apply these ideas in different fields of application.

The handout is thus an invitation to the bootstrap, showing its flexibility and its limitations. It is not a basic theoretical reference. More is available in the literature. It is mainly inspired from Efron and Tibshirani (1993) and Davison and Hinkley (1997). The presentation in Sections 1.1-1.3, follows some of the ideas in Politis (1995).

1.1 The Aim of Statistical Inference

• The Data Generating Process (DGP): \( X = (X_1, \ldots, X_n) \) is an i.i.d. sample from a population with c.d.f. \( F(x) = P(X_i \leq x) \); \( F \) is unknown (nonparametric case) or \( F(.) = F_\tau(.) \) where only \( \tau \in IR^k \) is unknown (parametric case).
• Quantity of interest $\theta(F)$ (mean, median, variance, ...). Examples:

\[ \theta_1(F) = \int x \, dF(x), \]
\[ \theta_2(F) = \int x^2 \, dF(x), \]
\[ \theta_3(F) = (1 - \alpha)\text{-quantile of } F(x) : \ P(X \leq \theta_3(F)) = 1 - \alpha \]

• A statistic $T(X)$ is an estimator of $\theta(F)$. Examples:

\[ T_1(X) = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \]
\[ T_2(X) = \overline{X^2} = \frac{1}{n} \sum_{i=1}^{n} X_i^2, \]
\[ T_3(X) = (1 - \alpha)\text{-quantile of } F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \]

• Measures of the accuracy of the estimator $T(X)$?

Quantities of interest:

\[ \text{Bias}_F(T) = E_F T(X) - \theta(F) \]
\[ \text{Var}_F(T) = E_F T^2(X) - (E_F T(X))^2 \]
\[ \text{MSE}_F(T) = E_F((T(X) - \theta(F))^2) \]
\[ G(x) = \text{Dist}_{T,F}(x) = P_F(T(X) \leq x) \]

• Often these quantities are unavailable ($F$ is unknown), sometimes some asymptotic approximations are available (as $n \to \infty$).
- **Example**: finding confidence intervals for $\theta(F)$: $(T(\mathcal{X}) - \theta(F))$ is called a *root*. Quantity of interest here:

$$H(x) = P_F(T(\mathcal{X}) - \theta(F) \leq x)$$

where

$$H(x) = Dist_{T,F}(x + \theta(F)) = G(x + \theta(F))$$

- If $H(x)$ is known, $(T(\mathcal{X}) - \theta(F))$ is called a **pivotal root**:

  - $P_F\left(u\left(\frac{\alpha}{2}\right) \leq T(\mathcal{X}) - \theta(F) \leq u\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha,$

  where $u(a) = H^{-1}(a)$ is the $a-$quantile of $H(\cdot)$, i.e.

  $$H(u(a)) = P_F(T(\mathcal{X}) - \theta(F) \leq u(a)) = a$$

  - An equal-tailed $(1 - \alpha)100\%$ confidence interval for $\theta(F)$ is then

    $$\left[T(\mathcal{X}) - u\left(1 - \frac{\alpha}{2}\right), T(\mathcal{X}) - u\left(\frac{\alpha}{2}\right)\right]$$

- If $H(x)$ is not known: useless.

- **Basic Question**: What are the sampling properties of such roots?

  - **Example 1**: estimating the mean in a **Normal** population:

    $$X \sim N(\mu, \sigma^2)$$

    - Possible roots:

      $$R1_n = \sqrt{n}(\bar{X} - \mu) \sim N(0, \sigma^2)$$

      $$R2_n = \sqrt{n}\frac{(\bar{X} - \mu)}{S} \sim t_{n-1}$$

    where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n}(X_i - \bar{X})^2$. 


– Confidence interval for $\mu$ using $R_{1n}$ if $\sigma$ known:

$$\left[ \bar{X} - z_{(1-\alpha/2)} \frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

where $z_{\alpha/2} = -z_{(1-\alpha/2)}$ is the $(\alpha/2)$-quantile of $N(0, 1)$.

– Confidence interval for $\mu$ using $R_{2n}$ if $\sigma$ unknown:

$$\left[ \bar{X} - t_{n-1,(1-\alpha/2)} \frac{S}{\sqrt{n}}, \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right]$$

where $t_{n-1,\alpha/2} = -t_{n-1,(1-\alpha/2)}$ is the $(\alpha/2)$-quantile of $t_{n-1}$.

– When $\sigma$ is unknown, $R_{1n}$ is a root and $R_{2n}$ is a pivotal root.

– Example 2: estimating the mean in an unknown population: $X \sim (\mu, \sigma^2)$ where both $(\mu, \sigma^2)$ are unknown

– Only asymptotic results

– $R_{1n}$ is a root with $R_{1n} \sim AN(0, \sigma^2)$

– $R_{2n}$ is an asymptotically pivotal root: $R_{2n} \sim AN(0, 1)$.

– Example 3: estimating the median of an unknown population: only few asymptotic results for some particular population.

– Example 4: nonparametric statistics, even more complicated.

• Problem: except for simple problems (where the root is pivotal), sampling distribution of roots are difficult to obtain and/or depend on $F$ which is unknown. Moreover, the results are generally asymptotic ones (approximations).
• The bootstrap provides easy-to-use and powerful methods for this purpose.
  - Provides approximations of sampling distribution of “any” root
    \( T(\mathcal{X} - \theta(F)) \) or of “any” statistics \( T(\mathcal{X}) \)
  - Even if asymptotic results are available, bootstrap generally provides better approximations.
  - In complicated problems, there are often no alternatives . . .
  - There are a few cases where the bootstrap does not work.

• The term **BOOTSTRAP** : Adventures of Baron Munchausen (R.E. Raspe 18th century) : “to pull oneself up by ones’ bootstrap”.

### 1.2 The Bootstrap Principle

#### 1.2.1 The Plug-in Principle

• The problem comes from the fact that \( F \) is unknown. We have a “good” estimator of \( F \) : \( F_n \) the empirical c.d.f. (putting a mass \( \frac{1}{n} \) at each \( X_i \)).

\[
F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)
\]

• The Plug-in Principle: replace \( F \) by \( F_n \) (an old idea !)

\[
\theta(F) = \int x dF(x)
\]

\[
T(\mathcal{X}) = \hat{\theta}(F) = \theta(F_n) = \int x dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}
\]
• The Bootstrap Principle:

- **Real world**: the DGP provides a sample \( \mathcal{X} = \{X_1, \ldots, X_n\} \) drawn from an unknown population with unknown c.d.f. \( F \).
  
  - A statistic \( T(\mathcal{X}) \) provide an estimator of the unknown \( \theta(F) \)
  
  - Sampling properties of \( T(\mathcal{X}) \) rely on the (often) unknown
    \[
    G(x) = \text{Dist}_{T,F}(x) = P_F(T(\mathcal{X}) \leq x).
    \]

- **Bootstrap world**: consider the DGP*, where a pseudo-sample \( \mathcal{X}^* = \{X_1^*, \ldots, X_n^*\} \) is drawn from the given population \( \mathcal{X} \) with known c.d.f. \( F_n \) (mass 1/n at each observed \( X_i \)).

  - conditionally to \( \mathcal{X} \), each observation \( X_i^* \) has a c.d.f.:
    \[
    P(X_i^* \leq x | \mathcal{X}) = F_n(x)
    \]
  
  - the function \( T(\cdot) \) applied to \( \mathcal{X}^* \) provide an estimator \( T(\mathcal{X}^*) \) of \( \theta(F_n) \), which, conditionally to \( \mathcal{X} \) is known
  
  - conditionally to \( \mathcal{X} \), the sampling properties of the estimator \( T(\mathcal{X}^*) \) depends only on \( F_n \), so are known (although they might be difficult to compute: see below).

- **The bootstrap idea**:

  The known sampling properties of \( T(\mathcal{X}^*) \) could mimic the (unknown) sampling properties of \( T(\mathcal{X}) \)
• Bootstrap estimates for the quantities of interest:

\[
\begin{align*}
\text{Bias}^* (T) & = \text{Bias}_{F_n} (T) = E_{F_n} T(X^*) - \theta(F_n) \\
\text{Var}^* (T) & = \text{Var}_{F_n} (T) = E_{F_n} T^2(X^*) - (E_{F_n} T(X^*))^2 \\
\text{MSE}^* (T) & = \text{MSE}_{F_n} (T) = E_{F_n} \left( [T(X^*) - \theta(F_n)]^2 \right) \\
\text{Dist}^*_T (x) & = \text{Dist}_{T,F_n} (x) = P_{F_n} (T(X^*) \leq x) = P \left( T(X^*) \leq x \mid X \right)
\end{align*}
\]

• In particular, \( \hat{G}(x) = \text{Dist}_{T,F_n} (x) \) is the bootstrap estimate of \( G(x) = \text{Dist}_{T,F} (x) = P_F (T(X) \leq x) \)

N.B.: the bootstrap estimate of \( H(x) = P_F (T(X) - \theta(F) \leq x) \) is given by

\[
\hat{H}(x) = P_{F_n} (T(X^*) - \theta(F_n) \leq x) \\
= P \left( T(X^*) - \theta(F_n) \leq x \mid X \right) \\
= \hat{G}(x + \theta(F_n))
\]

• The Key relation:

under “regularity conditions” (see below), when \( n \) is large:

\[
\hat{G}(x) = \text{Dist}_{T,F_n} (x) \approx \text{Dist}_{T,F} (x) = G(x)
\]
• Remarks

1. Often, but not necessarily, \( T(\mathcal{X}) = \theta(F_n) \). This may depend on the appropriate definition of \( \theta(F) \). More generally, \( T(\mathcal{X}) = g_n(\theta(F_n)) \), where \( g_n \) is known (example: \( S^2 = \frac{n}{n-1} \sigma^2(F_n) \)).

2. Parametric Bootstrap:

   - If \( F(.) = F_\tau(.) \) where only \( \tau \in \mathbb{R}^k \) is unknown.
   
   - Same approach but here \( F_n(.) = F_\hat{\tau}(.) \) where \( \hat{\tau} = \hat{\tau}(\mathcal{X}) \) is a (consistent) estimator of the parameter(s) \( \tau \)

3. In the bootstrap world, conditionally on \( \mathcal{X} \), everything is known but may be difficult to compute:

   \[ \Rightarrow \text{Monte-Carlo approximations} \]

The next (simple) example show a case where some bootstrap estimates can be computed analytically.

• Example: Nonparametric Bootstrap, Estimation of a mean \( \mu \) by \( \bar{X} \) where \( X \sim (\mu, \sigma^2) \).

   - Let \( (X_1^*, X_2^*, \ldots, X_n^*) \) be an i.i.d. sample drawn from the DGP*
   
   - We have in the real world: \( X \sim F(.) \)

\[ E_F(X) = \mu \]

\[ Var_F(X) = \sigma^2 \]
– We have in the bootstrap world: \( X^*|\mathcal{X} \sim F_n(\cdot) \)

\[
E^*(X^*) = E_{F_n}(X^*) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}
\]

\[
Var^*(X^*) = Var_{F_n}(X^*) = E_{F_n}(X^* - E_{F_n}(X^*))^2
= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n-1}{n} S^2
\]

– Sampling properties:

- Sample moments in the real world:

\[
E(\bar{X}) = E_F(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E_F(X) = \mu \text{ and } Bias(\bar{X}) = 0
\]

\[
Var(\bar{X}) = Var_F(\bar{X}) = \frac{1}{n} Var_F(X) = \frac{\sigma^2}{n}
\]

- Sample moments in the bootstrap world:

\[
E^*(\bar{X}^*) = E_{F_n}(\bar{X}^*) = \frac{1}{n} \sum_{i=1}^{n} E_{F_n}(X^*) = \bar{X}
\]

\[
Var^*(\bar{X}^*) = Var_{F_n}(\bar{X}^*) = \frac{1}{n} Var_{F_n}(X^*) = \frac{n-1}{n} \frac{S^2}{n}
\]

- Bootstrap estimates of \( Bias \) and of \( Var \) of \( \bar{X} \):

\[
Bias(\bar{X}) \approx \widehat{Bias}(\bar{X}) = Bias^*(\bar{X}^*) = E^*(\bar{X}^*) - \bar{X} = 0
\]

\[
Var(\bar{X}) \approx \widehat{Var}(\bar{X}) = Var^*(\bar{X}^*) = \frac{n-1}{n} \frac{S^2}{n}
\]
1.2.2 Computational aspects: Monte-Carlo Methods

• If \( F \) is known: Monte-Carlo approximations are available
  
  - In the simplest cases, \( \text{Bias}_F(T) \), \( \text{Var}_F(T) \), \( \text{MSE}_F(T) \) and \( \text{Dist}_{T,F}(x) \) could eventually be calculated by analytical methods.

  - Example: If \( X_i \sim \text{Expo}(\mu) \) with \( \mu = E(X_i) \) (see the Appendix A). We chose the statistic \( T(\mathcal{X}) = \bar{X} \):
    - then \( \bar{X} \sim \Gamma(n, \mu/n) \)
    \[
    G(x) = P_F(\bar{X} \leq x) = \text{Dist}_{T,F}(x) = \int_{0}^{x} \frac{u^{n-1} \exp(-nu/\mu)}{(\mu/n)^n \Gamma(n)} \, du
    \]
    - if \( n \) is large, then \( \bar{X} \sim \text{AN}(\mu, \mu^2/n) \)
    \[
    G(x) = P_F(\bar{X} \leq x) = \text{Dist}_{T,F}(x) \approx \Phi \left( \frac{x - \mu}{\mu/\sqrt{n}} \right)
    \]
  - \( G(x) \) can also be approximated by Monte-Carlo simulations:
    - Draw \( B \) samples \( (\mathcal{X}^{(b)}, b = 1, \ldots, B) \) from the known \( F \), (where \( B \) is large).
    - This provides \( B \) samples \( \mathcal{X}^{(b)} = (X_1^{(b)}, \ldots, X_n^{(b)}) \) where \( X_i^{(b)} \) is drawn independently from \( F \).
    - Compute \( T(\mathcal{X}^{(b)}), b = 1, \ldots, B \), (in our example, \( T(\mathcal{X}^{(b)}) = \bar{X}^{(b)} = (1/n) \sum_{i=1}^{n} X_i^{(b)} \)).
    - Monte-Carlo approximations: strong law of large numbers
      \[
      G(x) = \text{Dist}_{T,F}(x) \approx \frac{1}{B} \sum_{b=1}^{B} I(T(\mathcal{X}^{(b)}) \leq x)
      \]
The quality of the approximations depends on $B$ which can be set as large as we want (only computer limitations).

We have also the Monte-Carlo estimates of the $Bias$, $Var$ and $MSE$:

$$\text{Bias}_F(T) \approx \frac{1}{B} \sum_{b=1}^{B} T(X^{(b)}) - \theta(F)$$

$$\text{Var}_F(T) \approx \frac{1}{B} \sum_{b=1}^{B} T^2(X^{(b)}) - \left( \frac{1}{B} \sum_{b=1}^{B} T(X^{(b)}) \right)^2$$

$$\text{MSE}_F(T) \approx \frac{1}{B} \sum_{b=1}^{B} ((T(X^{(b)}) - \theta(F))^2)$$

where, since $F$ is known, $\theta(F)$ is also known.

Often useless because either we know the sampling distribution of $T(X)$ or we know at least an asymptotic approximation.

**Example: Illustration of the Monte-Carlo principle.**

- Exponential population: $X_i \sim \text{Expo } (\mu)$; $\mu = 50, n = 10$

  $\theta(F) = \mu$ and $T(X) = \theta(F_n) = \bar{X}$

- True sampling distribution : $\bar{X} \sim \Gamma(n, \frac{\mu}{n})$

- CLT approximation : $\bar{X} \sim AN(\mu, (\frac{\mu^2}{n}))$

- $B$ Monte-Carlo simulations: empirical distribution of $\bar{X}^{(b)}$, $b = 1, \ldots, B$: see Figure 1.1.
Figure 1.1: Sampling distribution of $\bar{X}$, $X_i \sim \text{Expo} (\mu = 50)$ with $n = 10$. True Gamma (dashed), Normal approximation (dashdot) and MC approximation (solid): top $B = 500$ and bottom $B = 5000$.

- Remark:

Always impossible to perform Monte-Carlo approximations in practical applications:

since we do not know $F$ !
• Sampling distributions in the Bootstrap world: $F_n$ is known!

**Monte-Carlo approximations are always available!**

- Draw $B$ samples $(\mathcal{X}^*(b), b = 1, \ldots, B)$ from $F_n$.
- This provides $B$ “pseudo-samples” $\mathcal{X}^*(b) = (X_1^*(b), \ldots, X_n^*(b))$
  where $X_i^*(b)$ is drawn independently from $F_n$,
- practically, $X_i^*(b)$ is drawn with replacement from $\mathcal{X} = (X_1, \ldots, X_n)$.
- Compute $T(\mathcal{X}^*(b)), b = 1, \ldots, B$.
- Monte-Carlo approximations of the bootstrap estimates:
  \[
  \text{Bias}^*(T) \approx \frac{1}{B} \sum_{b=1}^{B} T(\mathcal{X}^*(b)) - \theta(F_n)
  \]
  \[
  \text{Var}^*(T) \approx \frac{1}{B} \sum_{b=1}^{B} T^2(\mathcal{X}^*(b)) - \left( \frac{1}{B} \sum_{b=1}^{B} \frac{T(\mathcal{X}^*(b))}{B} \right)^2
  \]
  \[
  \text{MSE}^*(T) \approx \frac{1}{B} \sum_{b=1}^{B} \left( (T(\mathcal{X}^*(b)) - \theta(F_n))^2 \right)
  \]
  \[
  \hat{G}(x) = \text{Dist}^*_T(x) \approx \frac{1}{B} \sum_{b=1}^{B} I(T(\mathcal{X}^*(b)) \leq x)
  \]

• The bootstrap principle is the plug-in principle and NOT the simulation principle which is only a way for computing the bootstrap estimates (See Figure 1.2: the bootstrap analogy).
• Remark: parametric bootstrap, the same but here $F_n = F_\hat{\tau}$, the $X_i^*(b)$ are independently drawn from the cdf $F_\hat{\tau}$.
  Example: If $X \sim Expo(\mu)$ then $X^* \sim Expo(\bar{x})$.
Figure 1.2: The bootstrap analogy: the russian dolls (from Hall, 1992).
**Example:** Nonparametric bootstrap $X \sim (\mu, \sigma^2)$, $n = 5$

The DGP is completely unknown: we observe the data,

$$(X_1, X_2, X_3, X_4, X_5) = (22, 10, 15, 18, 20)$$

<table>
<thead>
<tr>
<th>Original data vector</th>
<th>$\bar{x}$</th>
<th>$S^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 10 15 18 20</td>
<td>17.00</td>
<td>22.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bootstrap data vectors</th>
<th>$\bar{x}_b$ for $b=1,10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 10 18 18 20</td>
<td>15.20</td>
</tr>
<tr>
<td>10 20 10 18 20</td>
<td>15.60</td>
</tr>
<tr>
<td>20 22 10 22 15</td>
<td>17.80</td>
</tr>
<tr>
<td>15 22 22 15 15</td>
<td>17.80</td>
</tr>
<tr>
<td>20 10 15 22 15</td>
<td>16.40</td>
</tr>
<tr>
<td>20 20 20 18 10</td>
<td>17.60</td>
</tr>
<tr>
<td>15 18 15 18 15</td>
<td>16.20</td>
</tr>
<tr>
<td>15 22 18 20 15</td>
<td>18.00</td>
</tr>
<tr>
<td>20 20 18 15 18</td>
<td>18.20</td>
</tr>
<tr>
<td>22 20 20 18 10</td>
<td>18.00</td>
</tr>
</tbody>
</table>

| Bootstrap mean and bootstrap variance of $\bar{x}_b$ for $B=1000$ |
|-------------------------|-------------------------|
| 17.0504                 | 3.3949                  |

Sampling distribution: see figure 1.3.

![Sampling distribution of the bootstrap values of $\bar{x}$](image)

Figure 1.3: $\hat{G}(x)$, the bootstrap sampling distribution of $\bar{X}, n = 5$ and $B = 1000$. 
• **Example: Bootstrapping the correlation**

Let a sample be generated from a bivariate Normal distribution

\[ X \sim N_2 \left( \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} 4 & 1.5 \\ 1.5 & 1 \end{array} \right) \right) \]

Here \( \rho = 0.75 \). The MLE estimator is \( \hat{\rho}(\mathbf{X}) = r \), the empirical correlation.

- We want the **sampling distribution of** \( \hat{\rho}(\mathbf{X}) \)

- Let a sample of size \( n = 100 \) giving \( \hat{\mu} = (-0.1077 - 0.0908)' \) and \( S = \left( \begin{array}{cc} 4.2567 & 1.3662 \\ 1.3662 & 0.9974 \end{array} \right) \). So, \( r = 0.6631 \).

- **No real alternatives** here to estimate the sampling distribution of \( \hat{\rho}(\mathbf{X}) \)

  (N.B. some asymptotic results exist leading to Normal distributions, using Fisher information matrix or, nonparametric delta methods, ...).

- We obtain **the bootstrap approximations** of the unknowns \( E(\hat{\rho}) \) and \( Var(\hat{\rho}) \):

\[
E^*(\hat{\rho}) = \frac{1}{5000} \sum_{b=1}^{5000} \hat{\rho}(\mathbf{X}^{*,b}) = 0.6622
\]

\[
Var^*(\hat{\rho}) = \left( \frac{1}{5000} \sum_{b=1}^{5000} \hat{\rho}^2(\mathbf{X}^{*,b}) \right)^2 - \left( \frac{1}{5000} \sum_{b=1}^{5000} \hat{\rho}(\mathbf{X}^{*,b}) \right)^2 = 0.0023.
\]
– Figure 1.4 displays the Bootstrap distribution of $\hat{\rho}(\mathcal{X}^*)$

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.4}
\caption{Nonparametric Bootstrap approximation, $B = 5000$ of the sampling distribution of $\hat{\rho}(\mathcal{X})$. Left panel: histogram, right panel smoothed density estimate.}
\end{figure}

• Example: Duration times data: $n = 10$ observations of duration times (failures,...)

$$\mathcal{X} = (X_1, \ldots, X_{10}) = (1 \ 5 \ 12 \ 15 \ 20 \ 26 \ 78 \ 145 \ 158 \ 358)$$

The exponential model might be reasonable for $X$: $X \sim Expo(\mu)$ where $\mu = E[X]$.

Figure 1.5 shows that the exponential model might be reasonable but no enough observations ($n = 10$) to be sure. We have $\bar{x} = 81.80$ and $s = 112.94$ (Note the small overdispersion w.r.t. exponential assumption: $\mu = \sigma$).
We are interested in the sampling distribution of $T(\mathcal{X}) = \bar{X}$, as potential estimator of $\mu$. We have $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \sigma^2/n$. We are offered different approaches.

1. Parametric analysis small sample: if we believe in the exponential model, we have the exact result: $\bar{X} \sim \Gamma(n, \mu/n)$. This can be estimated by

$$\bar{X} \sim \Gamma(n, \bar{x}/n)$$

This kind of result is not always available (if more complicated statistics and/or models)!

2. Parametric case, asymptotic result: if we believe in the exponential model, $\sigma = \mu$, so that $\bar{X} \sim N(\mu, \mu^2/n)$. This can be
estimated by

$$\bar{X} \sim N(\bar{x}, \frac{\bar{x}^2}{n})$$

This kind of result is not always available (if more complicated statistics and/or models)!

3. Parametric bootstrap: if we believe in the exponential model, we simulate $B = 5000$ samples $X^{\ast(b)}$ of size $n = 10$ drawn from $Expo(\bar{x})$ and compute for each $\bar{X}^{\ast(b)}$. The 5000 values $\bar{X}^{\ast(b)}, b = 1,5000$ provide the empirical parametric bootstrap approximation of the sampling distribution of $\bar{X}$. This result is always available.

4. Non-Parametric asymptotic result: we do not believe in the exponential model, so that $\bar{X} \sim N(\mu, \sigma^2/n)$. This can be estimated by

$$\bar{X} \sim N(\bar{x}, \frac{s^2}{n})$$

This kind of result is not always available (if more complicated statistics and/or models)!

5. Non-Parametric bootstrap: we simulate $B = 5000$ samples $X^{\ast(b)}$ of size $n = 10$ drawn from $X$ and compute for each $\bar{X}^{\ast(b)}$. The 5000 values $\bar{X}^{\ast(b)}, b = 1,5000$ provide the empirical bootstrap approximation of the sampling distribution of $\bar{X}$. This result is always available.

Figure 1.6 shows the 5 different corresponding densities.
Figure 1.6: Various estimates of the sampling distributions of $\bar{X}$. Dotted (big dots) is the assumed parametric density (gamma), Dash-dot is the parametric normal approximation, Solid is the parametric bootstrap approximation, Dotted (small dots) is the nonparametric normal approximation and Dashed is the nonparametric bootstrap approximation.
• Remark 1:

- If the Exponential is “true”, the 3 distributions (Gamma, parametric bootstrap and parametric asymptotic normal) are similar. The bootstrap here is much better than normal approximation.

- The two nonparametric approaches give different results! (Exponential hypothesis is questionable). The nonparametric normal gives some mass below zero (which is impossible...). We will see later the bootstrap is approximation should be slightly better.

• Remark 2:

Each sampling distribution provides estimates of $E[\bar{X}]$ and of $Std[\bar{X}]$

1. small sample parametric:

$\hat{E}[\bar{X}] = \bar{x} = 81.80$ and $\hat{Std}[\bar{X}] = \bar{x}/\sqrt{n} = 25.87$

2. asymptotic parametric:

$\hat{E}[\bar{X}] = \bar{x} = 81.80$ and $\hat{Std}[\bar{X}] = \bar{x}/\sqrt{n} = 25.87$

3. bootstrap parametric:

$\hat{E}[\bar{X}] \approx 82.15$ and $\hat{Std}[\bar{X}] \approx 26.04$

4. asymptotic Non-parametric:

$\hat{E}[\bar{X}] = \bar{x} = 81.80$ and $\hat{Std}[\bar{X}] = s/\sqrt{n} = 35.71$

5. bootstrap Non-parametric:

$\hat{E}[\bar{X}] \approx 82.52$ and $\hat{Std}[\bar{X}] \approx 33.74$
• A Simulated Example

Bootstrap approximation for the sampling distribution of $\bar{X}$ with one simulated sample $X_i \sim \text{Expo} (\mu = 50)$, see Figure 1.7.

Figure 1.7: Sampling distribution of $\bar{X}$ where $X_i \sim \text{Expo} (\mu = 50)$. Dashed line is the true Gamma $\Gamma(n, \mu/n)$, dot is the estimated Gamma $\Gamma(n, \bar{X}/n)$, dash-dot is the estimated Normal approximation $N(\bar{X}, s^2/n)$ and solid is the Nonparametric Bootstrap approximation. Here, $B = 5000$, top with $n = 10$ and bottom with $n = 100$. 
1.3 An Alternative to the Bootstrap: Asymptotic Theory

1.3.1 The Parametric Delta Method

- Asymptotics for \( T = T(\mathbf{X}) \)?
  - There are other methods for approximating the sampling distribution of a statistic \( T \). Most are based on the CLT. This allows, in particular to obtain asymptotic expression for the variance of \( T \) (could be useful for the Bootstrap-t method below).
  - Often \( T = T(\mathbf{X}) \) can be represented in terms of simpler statistics \( U_j = U_j(\mathbf{X}), j = 1, \ldots, p \), such as sample moments, for which asymptotic sampling distribution can be derived by using the CLT.

**Theorem 1.1. Delta Method.** Let \( \sqrt{n}(U - \mu) \sim \mathcal{N}_p(0, \Sigma) \) where \( \Sigma > 0 \) and \( T = g(U) \) where \( g : \mathbb{R}^p \rightarrow \mathbb{R}^q \), \( q \leq p \) and \( g \) are differentiable at \( u = \mu \). Then

\[
\sqrt{n}(T - g(\mu)) \sim \mathcal{N}_q(0, \Delta' \Sigma \Delta)
\]

where

\[
\Delta = \left. \frac{\partial g'(u)}{\partial u} \right|_{u=\mu}
\]
Sketch of the proof in the univariate case:

− it comes from Taylor series expansion:

\[ T = g(U) = g(\mu) + (U - \mu)\dot{g}(\mu) + o_p(n^{-1/2}), \]

where \( \dot{g}(\mu) = \partial g(u)/\partial u|_{u=\mu}. \)

− Then we have:

\[ T = g(\mu) - n^{-1/2}\dot{g}(\mu) \sigma Z + o_p(n^{-1/2}), \]

where \( Z \sim N(0, 1) \) and \( \sigma^2 = \Sigma = n \text{Var}(U). \)

− This provides also the asymptotic mean and variance of \( T. \)

**Example:**

− In the exponential model we have

\( \bar{X} \sim AN(\mu, \mu^2/n). \)

Let \( T = \log(\bar{X}) \) we obtain:

\[ T \sim AN(\log(\mu), 1/n). \]

− More generally \( \bar{X} \sim AN(\mu, \sigma^2/n) \) so that

\[ T = \log(\bar{X}) \sim AN(\log(\mu), \sigma^2/(n\mu^2)). \]
1.3.2 Nonparametric Delta Method, Influence Functions

- Let $T(\mathcal{X}) = \theta(F_n)$ be a plug-in estimator of $\theta(F)$. Let $G$ be a c.d.f. Under regularity condition on the function $\theta(\cdot)$ (It should be Gâteaux differentiable at $F$), the extension of Taylor expansion in functional spaces, can be written as (see Serfling, 1980):

$$
\theta(G) = \theta(F) + \int L_\theta(x; F) \, dG(x) + o(||G - F||),
$$

where $L_\theta$ is the Frechet first derivative of $\theta(\cdot)$ at $F$.

- If $H_x$ denote the c.d.f. corresponding to the Dirac measure at the point $x$, we have:

$$
L_\theta(x; F) = \lim_{\varepsilon \to 0} \frac{\theta((1 - \varepsilon)F + \varepsilon H_x) - \theta(F)}{\varepsilon} = \frac{\partial \theta((1 - \varepsilon)F + \varepsilon H_x)}{\partial \varepsilon} \bigg|_{\varepsilon=0}.
$$

- $L_\theta(x; F)$ is called the influence function of $\theta$ and its empirical approximation $L_\theta(x; F_n)$ is the empirical influence function.

- It can be shown that (let $G = F$ in Taylor expansion above):

$$
E_F(L_\theta(X; F)) = \int L_\theta(x; F) \, dF(x) = 0.
$$
– The nonparametric delta method is the application of the approximation with $G = F_n$:

$$\theta(F_n) = \theta(F) + \int L_\theta(x; F) \, dF_n(x) + o_p(||F_n - F||)$$

$$\approx \theta(F) + \int L_\theta(x; F) \, dF_n(x)$$

$$= \theta(F) + \frac{1}{n} \sum_{i=1}^{n} L_\theta(x_i; F).$$

– The CLT applied to the sum provides:

$$\theta(F_n) - \theta(F) \sim AN(0, var_L(F)).$$

- We have:

$$E(L_\theta(X; F)) = \int L_\theta(x; F) \, dF(x) = 0$$

$$var_L(F) = \frac{1}{n} Var(L_\theta(X; F)) = \frac{1}{n} \int L_\theta^2(x; F) \, dF(x).$$

– In practice, the latter is approximated by plugging $F_n$ in place of $F$:

$$var_L(F_n) = \frac{1}{n^2} \sum_{i=1}^{n} L_\theta^2(x_i; F_n)$$

Note that we have also for the empirical influence functions:

$$\int L_\theta(x; F_n) \, dF_n(x) = \sum_{i=1}^{n} L_\theta(x_i; F_n) = 0$$
Example

- **Sample mean:** $\theta(F) = \int xdF(x)$ and $\theta(F_n) = \bar{X}$. We have

  - $\theta ((1 - \varepsilon)F + \varepsilon H_x) = (1 - \varepsilon)\mu + \varepsilon x$ so that
    
    $$L_{\theta}(x; F) = \frac{\partial ((1 - \varepsilon)\mu + \varepsilon x)}{\partial \varepsilon} \bigg|_{\varepsilon=0} = x - \mu.$$  

  - Now $L_{\theta}(x; F_n) = x - \bar{x}$ and $L_{\theta}(x_i; F_n) = x_i - \bar{x}$.

  - Finally:
    
    $$\text{var}_L(F_n) = \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n-1}{n^2} s^2 = \frac{v}{n}.$$  

      where $v$ is the (plug-in) empirical variance.

- So that the nonparametric delta method provides

  $$\bar{X} - \mu \sim \text{AN}(0, \frac{v}{n}).$$

- **Linear functionals:** if $\theta(F) = \int g(x)dF(x)$ the same argument as for the mean gives

  $$L_{\theta}(x; F) = g(x) - \theta(F) \Rightarrow L_{\theta}(x_i; F_n) = g(x_i) - \theta(F_n) \text{ etc.}$$

Examples: all moments of $X$ about zero: $g(x) = x^k$. 

- **Function of simple functionals:** if \( \theta(F) = g(\theta_1(F), \ldots, \theta_p(F)) \), we have (by the chain rule)

\[
L_\theta(x; F) = \sum_{j=1}^{p} \frac{\partial g}{\partial \theta_j} L_{\theta_j}(x; F).
\]

- **Correlation coefficient:** This is a function of sample moments (see Davison and Hinkley, 1997, example 2.18)

\[
\rho(F) = \frac{\mu_{11} - \mu_{10} \mu_{01}}{\sqrt{(\mu_{20} - \mu_{10}^2)(\mu_{02} - \mu_{01}^2)}},
\]

where \( \mu_{rs} = E[X^rY^s] \). The influence function turns out to be:

\[
L_\rho(x, y; F) = x_s y_s - \frac{1}{2} \rho(x_s^2 + y_s^2),
\]

where \( x_s \) and \( y_s \) are the centered and reduced versions of \( x \) and \( y \) respectively.

- **Link with the Jackknife:** the Jacknife can be used to approximate \( L_\theta(x_i; F_n) \) (see Chapter 7).
1.4 Bootstrap Estimation of Bias and Standard Errors

1.4.1 Plug-in principle

- We want to estimate the bias and standard error of $T(\mathcal{X})$ as estimator of $\theta(F)$, where $T(\mathcal{X})$ may or may not be the plug-in estimator $\theta(F_n)$. The bias of $T$ is given by

$$Bias(T) = E_F[T(\mathcal{X})] - \theta(F)$$

The variance is

$$Var(T) = E_F[T^2(\mathcal{X})] - (E_F[T(\mathcal{X})])^2$$

- The bootstrap estimation of the bias is thus given by

$$Bias^*(T) = E_{F_n}[T(\mathcal{X}^*)] - \theta(F_n),$$

where $\theta(F_n)$ is the plug-in version of $\theta(F)$. This is approximated by:

$$Bias^*(T) \approx \frac{1}{B} \sum_{b=1}^B T(\mathcal{X}^*(b)) - \theta(F_n)$$

- The bootstrap estimation of the variance is:

$$Var^*(T) = E_{F_n}[T^2(\mathcal{X}^*)] - (E_{F_n}[T(\mathcal{X}^*)])^2,$$

approximated by:

$$Var^*(T) \approx \frac{1}{B} \sum_{b=1}^B T^2(\mathcal{X}^*(b)) - \left( \frac{1}{B} \sum_{b=1}^B T(\mathcal{X}^*(b)) \right)^2.$$

- The bootstrap estimation of the standard error is thus:

$$Std^*(T) = \sqrt{Var^*(T)}$$
• How to choose $B$ to obtain a good approximation of the $Bias^*(T)$?
  
  – Denote by $Bias_B^*(T)$ the approximation and by $Bias_\infty^*(T)$ the true value of the estimate. By the CLT Theorem we have

  $$\text{Prob}_{F_n} \left\{ \left| \frac{1}{B} \sum_{b=1}^B T(X^*(b)) - E_{F_n}[T(X^*)] \right| \leq 2 \frac{\text{Std}^*(T)}{\sqrt{B}} \right\} \approx 0.95$$

  Equivalently:

  $$\text{Prob}_{F_n} \left\{ |Bias_B^*(T) - Bias_\infty^*(T)| \leq 2 \frac{\text{Std}^*(T)}{\sqrt{B}} \right\} \approx 0.95$$

  The “Error Bound” $2 \text{Std}^*(T)/\sqrt{B}$ gives a rough bound to the error in estimating the bias by $Bias_B^*(T)$.

• Examples

  1. **Ratio of two means:** $\theta = \mu_X/\mu_Y$, two independent samples of size $n = 10$. We obtain $T = \hat{\theta} = 2.8574$.

     | $x$     | $y$     |
    |---------|---------|
    | 137.1112| 7.1518  |
    | 380.0981| 1.6332  |
    | 205.2985| 16.4597 |
    | 2.3561  | 0.6340  |
    | 94.6776 | 25.7834 |
    | 61.1465 | 36.0518 |
    | 145.5394| 16.7388 |
    | 70.8571 | 88.8102 |
    | 173.8492| 54.0011 |
    | 7.4148  | 200.1222|

    -------------------------
    | $xbar$  | $ybar$  |
    | 127.8348| 44.7386 |
    | $s_x$   | $s_y$   |
    | 111.1522| 60.9250 |
Table 1.1: Choice of $B$ for Bias estimation of $\mu_X/\mu_Y$.

<table>
<thead>
<tr>
<th>Value of $B$</th>
<th>$Bias^*_B(T)$</th>
<th>$Std^*_B(T)$</th>
<th>error bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.5877</td>
<td>2.3578</td>
<td>0.4716</td>
</tr>
<tr>
<td>200</td>
<td>0.4843</td>
<td>1.7481</td>
<td>0.2472</td>
</tr>
<tr>
<td>400</td>
<td>0.7039</td>
<td>2.1431</td>
<td>0.2143</td>
</tr>
<tr>
<td>800</td>
<td>0.5994</td>
<td>2.2919</td>
<td>0.1621</td>
</tr>
<tr>
<td>1000</td>
<td>0.6711</td>
<td>2.2518</td>
<td>0.1424</td>
</tr>
<tr>
<td>2000</td>
<td>0.6799</td>
<td>2.2247</td>
<td>0.0995</td>
</tr>
<tr>
<td>5000</td>
<td>0.6247</td>
<td>2.0510</td>
<td>0.0580</td>
</tr>
<tr>
<td>10000</td>
<td>0.6109</td>
<td>2.0707</td>
<td>0.0414</td>
</tr>
</tbody>
</table>

See Table 1.1: for estimating the bias with a precision of, say 10%, we need more than $B = 5000$ replications. Then, at the level 95%, the $Bias^*_B(T)$ is bounded by $0.6247 + 0.0580 = 0.6827$.

NB: the two samples were generated from two independent exponentials with mean $\mu_X = 100, \mu_Y = 50$, so $\theta = 2$.

2. Estimation of the log of a mean: $\theta = \log(\mu)$ with the estimator $T = \log(\bar{x})$ with a sample of size $n = 8$. The sample is

$x=(28.7161, 36.2561, 40.2591, 0.8056, 88.3621, 157.1984, 17.5921, 1.8090)$

$xbar=46.3748$ and $s_X=52.6274$.

The estimator turns out to be $T = \hat{\theta} = \log(\bar{x}) = 3.8368$. The results are reported in Table 1.2, with the same conclusion. Here, for $B = 5000$, the negative bias estimate is bounded by $-0.0845$.

NB: the sample was generated from an exponential with mean $\mu_X = 100$, so that $\theta = 4.6052$ and $AStd(\hat{\theta}) = 1/\sqrt{8} = 0.3536$.  

1.4.2 Correction for the bias

- If $T(\mathcal{X})$ is biased, it is tempting to correct for the bias.

  - Usually

    $T_{\text{corr}}(\mathcal{X}) = T(\mathcal{X}) - \text{Bias}^*(T)$

    $= 2T(\mathcal{X}) - \frac{1}{B} \sum_{b=1}^{B} T(\mathcal{X}^{*(b)})$

- This can be dangerous in practice: due to the additional noise in $T_{\text{corr}}(\mathcal{X})$ due to the variance of $\text{Bias}^*(T)$: $MSE_F(T_{\text{corr}}(\mathcal{X}))$ could be greater than $MSE_F(T(\mathcal{X}))$.

**Table 1.2:** Choice of $B$ for Bias estimation for estimating $\log(\mu)$.

<table>
<thead>
<tr>
<th>Value of $B$</th>
<th>$\text{Bias}^*_B(T)$</th>
<th>$\text{Std}^*_B(T)$</th>
<th>error bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>-0.0680</td>
<td>0.4069</td>
<td>0.0814</td>
</tr>
<tr>
<td>200</td>
<td>-0.0751</td>
<td>0.4052</td>
<td>0.0573</td>
</tr>
<tr>
<td>400</td>
<td>-0.0814</td>
<td>0.3984</td>
<td>0.0398</td>
</tr>
<tr>
<td>800</td>
<td>-0.0949</td>
<td>0.4310</td>
<td>0.0305</td>
</tr>
<tr>
<td>1000</td>
<td>-0.0543</td>
<td>0.4050</td>
<td>0.0256</td>
</tr>
<tr>
<td>2000</td>
<td>-0.0702</td>
<td>0.4075</td>
<td>0.0182</td>
</tr>
<tr>
<td>5000</td>
<td>-0.0750</td>
<td>0.4119</td>
<td>0.0117</td>
</tr>
<tr>
<td>10000</td>
<td>-0.0762</td>
<td>0.4140</td>
<td>0.0083</td>
</tr>
</tbody>
</table>
– How to compute $MSE_F$?

- we have

$$Var_F(T_{corr}(X)) = Var_F \left( 2T(X) - \frac{1}{B} \sum_{b=1}^{B} T(X^{*\{b\}}) \right)$$

$$= 4Var_F(T(X)) + Var_F \left( \frac{1}{B} \sum_{b=1}^{B} T(X^{*\{b\}}) \right)$$

$$- 2 Cov_F \left( 2T(X), \frac{1}{B} \sum_{b=1}^{B} T(X^{*\{b\}}) \right)$$

$$= 4Var_F(T(X)) + Var_F[E^*(T(X^*))]$$

$$- 4 Cov_F[(T(X), E^*(T(X^*)) + O(B^{-1})$$

which contains many unknown quantities.

- $MSE_F(T(X))$ can be estimated by the bootstrap

- $MSE_F(T_{corr}(X))$ could be estimated by a double bootstrap (nested bootstrap: see below), in order to estimate $Var^*(T_{corr}(X))$ and $Bias^*(T_{corr}(X))$.

– In practice, correction is not recommended unless

$$Bias^*(T) \gg Std^*_B(T)$$

and even in this case, it should be better to try to find an better estimates with lower bias at the start.
- Efron and Tibshirani (1993) recommend, as a rule of thumb:

\[
\text{If } \frac{|Bias^*(T)|}{\text{Std}^*_B(T)} \leq 0.25 \text{ don’t correct for the bias},
\]

because in this case, the root mean squared error of \( T \) can not be larger than about 3.1\% of the standard error:

\[
RMSE_F(T) = \text{Std}_F(T) \sqrt{1 + \left( \frac{Bias_F(T)}{\text{Std}_F(T)} \right)^2}
\]

\[
\approx \text{Std}_F(T) \left\{ 1 + \frac{1}{2} \left( \frac{Bias_F(T)}{\text{Std}_F(T)} \right)^2 \right\}
\]

- In the examples above (with \( B = 5000 \)):

- For the estimation of the ratio of means we have

\[
|Bias^*(T)|/\text{Std}^*_B(T) = 0.6247/2.0510 = 0.3046
\]

\[
\hat{\theta}_{corr} = \hat{\theta} - Bias^*(T) = 2.8574 - 0.6247 = 2.2327
\]

- For the estimation of \( \log(\mu) \) we have

\[
|Bias^*(T)|/\text{Std}^*_B(T) = 0.0750/0.4119 = 0.1821
\]

\[
\hat{\theta}_{corr} = \hat{\theta} - Bias^*(T) = 3.8368 - (-0.0750) = 3.9118
\]

but here, the latter correction is certainly not recommended.
• Monte-Carlo evaluation of BIAS and MSE

- Generate samples of size \( n \) from known DGP (\( F \) and \( \theta = \theta_0 \) is known). For each simulated sample \( m = 1, \ldots, M \), compute \( \hat{\theta}_m \) and \( \hat{\theta}_{corr,m} \) (by bootstrap), we have

\[
\text{BIAS}(\hat{\theta}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_m - \theta_0)
\]

\[
\text{MSE}(\hat{\theta}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_m - \theta_0)^2
\]

\[
\text{BIAS}(\hat{\theta}_{corr}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_{corr,m} - \theta_0)
\]

\[
\text{MSE}(\hat{\theta}_{corr}) \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{\theta}_{corr,m} - \theta_0)^2
\]

- Example 1: simulation of \( M = 1000 \) random samples of size \( n = 10 \) of pairs \((X_i, Y_i)\) where \( X \sim \text{Exp}(\mu_X = 100) \) and \( Y \sim \text{Exp}(\mu_Y = 50), X \) independent of \( Y \). We obtain with \( B = 5000 \) (comp. time = 352”, pentium M, 1.8 Ghz), \( \theta_0 = \mu_X / \mu_Y = 2 \)

\[
\text{BIAS}(\hat{\theta}) = 0.2175, \quad \text{MSE}(\hat{\theta}) = 1.2558
\]

\[
\text{BIAS}(\hat{\theta}_{corr}) = -0.0014, \quad \text{MSE}(\hat{\theta}_{corr}) = 0.9871.
\]
Example 2: simulation of $M = 1000$ random samples of size $n = 8$ from $X \sim \text{Exp}(\mu_X = 100)$. We obtain with $B = 5000$ (comp. time = 223", pentium M, 1.8 Ghz), $\theta_0 = \log(\mu) = 4.6052$:

\[
\text{BIAS}(\hat{\theta}) = -0.0631, \quad \text{MSE}(\hat{\theta}) = 0.1377
\]

\[
\text{BIAS}(\hat{\theta}_{\text{corr}}) = -0.0099, \quad \text{MSE}(\hat{\theta}_{\text{corr}}) = 0.1358.
\]

\%
MATLAB code
\n```
n=8; a=1; beta=100; mu=a*beta; ttrue=log(mu);
t0=clock; MC=1000; B=5000;
MC_that=[]; MC_tcorr=[];

for mc=1:MC
    x=gamrnd(a,beta,n,1);
m=mean(x); t=log(m);
NParT=[];
for bb=1:B
    xb=boot(x); mb=mean(xb); tb=log(mb);
    NParT=[NParT;tb];
end
mstar=mean(NParT); bias=mstar-log(m); tcorr=t-bias;
MC_that=[MC_that;t]; MC_tcorr=[MC_tcorr;tcorr];
end

TCPU=etime(clock,t0);
fprintf(' Elapsed CPU time for the MONTE-CARLO :%15.4f \n',TCPU)

BIAS_t=mean(MC_that)-ttrue;
BIAS_tcorr=mean(MC_tcorr)-ttrue;
MSE_t=mean((MC_that-ttrue*ones(MC,1)).^2);
MSE_tcorr=mean((MC_tcorr-ttrue*ones(MC,1)).^2);
```
1.5 Bootstrap Confidence Intervals

1.5.1 Basic bootstrap confidence intervals

- **Pivotal Roots**: Confidence Interval for $\theta(F)$ can be obtained from $H(x)$, the c.d.f. of the root $T(\mathcal{X}) - \theta(F)$:

  $\quad H(x) = P_F(T(\mathcal{X}) - \theta(F) \leq x)$

  - We have:

  $\quad P_F \left( u\left( \frac{\alpha}{2} \right) \leq T(\mathcal{X}) - \theta(F) \leq u\left( 1 - \frac{\alpha}{2} \right) \right) = 1 - \alpha$

  where $u(a) = H^{-1}(a)$ is the $a-$quantile of $H$, i.e.

  $\quad H(u(a)) = P_F(T(\mathcal{X}) - \theta(F) \leq u(a)) = a$

  - An equal-tailed $(1 - \alpha)100\%$ confidence interval for $\theta(F)$ is then

    $\left[T(\mathcal{X}) - u\left(1 - \frac{\alpha}{2}\right), \; T(\mathcal{X}) - u\left(\frac{\alpha}{2}\right)\right]$

  - Pivotal roots are needed to be able to find the quantiles $u(.)$: $H(x)$ has to be known.

- **Asymptotic Pivotal Roots**

  - If $T$ is the MLE:

    - Let $\ell(\mathcal{X}; \theta)$ be the log-likelihood function, and $\dot{\ell}(\mathcal{X}; \theta)$ be the score function, then (under regularity conditions):

      $\quad T(\mathcal{X}) - \theta(F) \sim AN \left( 0, \mathcal{F}_n^{-1}(\theta) \right)$,

    where $\mathcal{F}_n$ is the Fisher information.

      $\quad \mathcal{F}_n(\theta) = Var(\dot{\ell}(\mathcal{X}; \theta)) = -E \left[ \ddot{\ell}(\mathcal{X}; \theta) \right]$. 
- $\mathcal{F}_n(\theta)$ can be estimated by $\mathcal{F}_n(\hat{\theta})$, or more simply by its consistent estimator, the observed Fisher information $-\ell(\mathcal{X}; \hat{\theta})$.

- An equal-tailed $(1 - \alpha)100\%$ asymptotic confidence interval for $\theta(\mathcal{F})$ is then

$$\left[ T(\mathcal{X}) - z(1 - \frac{\alpha}{2})\hat{\sigma}, T(\mathcal{X}) - z(\frac{\alpha}{2})\hat{\sigma} \right]$$

where $\hat{\sigma}^2 = AV ar(T(\mathcal{X}))$.

- More generally, if $T$ is the plug-in estimator, or a function of it, the Delta method could provide the estimates of the asymptotic variance of $T(\mathcal{X})$ (with the empirical influence function).

- “Hybrid Approach”: We can also use the (eventual) asymptotic normality, but the variance of $T(\mathcal{X})$, is obtained by the bootstrap approximation:

1. provide the bootstrap estimate of the sampling distribution of $T(\mathcal{X})$: $\hat{G}(x) = Dist^*_T(x)$;

2. by Q-Q plot (or normal probability plot), check for normality;

3. compute $Std^*(T)$;

4. the equal-tailed $(1 - \alpha)100\%$ confidence interval for $\theta(\mathcal{F})$ is then

$$\left[ T(\mathcal{X}) - z(1 - \frac{\alpha}{2})Std^*(T), T(\mathcal{X}) - z(\frac{\alpha}{2})Std^*(T) \right],$$

where $z_a$ is the $a$-quantile of a $N(0, 1)$ where by symmetry $z(\alpha/2) = -z(1 - \alpha/2)$. 
• The basic bootstrap confidence interval for \( \theta(F) \) is when we want to avoid the Normal approximation by using the bootstrap estimates \( \hat{G}(x) \) of the sampling distribution of \( T(\mathcal{X}) \), in place of the Normal, for estimating the unknown quantiles \( u(a) \).

The resulting confidence interval is:

\[
\left[ T(\mathcal{X}) - u^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) - u^*(\frac{\alpha}{2}) \right]
\]

where \( u^*(a) \) is the \( a \)-quantile of \( \hat{H}(x) = \hat{G}(x+\theta(F_n)) \), the bootstrap estimate of \( H(x) \):

\[
\hat{H}(x) = P_{F_n} (T(\mathcal{X}^*) - \theta(F_n) \leq x) = P(T(\mathcal{X}^*) - \theta(F_n) \leq x | \mathcal{X}) \\
\approx \frac{1}{B} \sum_{b=1}^{B} I \left( T(\mathcal{X}^{*(b)}) - \theta(F_n) \leq x \right).
\]

1.5.2 Practical computation of the quantiles \( u^*(a) \)

• The quantiles of \( \hat{H}(x) \) are more easily obtained from the quantiles \( v^*(a) \) of \( \hat{G}(x) = Dist_T(x) \):

\[
\hat{G}(x) = Dist_T^*(x) \approx \frac{1}{B} \sum_{b=1}^{B} I(T(\mathcal{X}^{*(b)}) \leq x).
\]

- Here \( v^*(a) = \hat{G}^{-1}(a) \), so,

\[
P_{F_n}(T(\mathcal{X}^*) \leq v^*(a)) = \hat{G}(v^*(a)) = a
\]
- Since $\hat{H}(x) = \hat{G}(x + \theta(F_n))$ we just have to shift the obtained quantiles to get $u^*(a)$:

$$u^*(a) = v^*(a) - \theta(F_n)$$

- The quantiles $v^*(a)$ of $\hat{G}(x) = Dist^*_T(x)$ are directly obtained from the following bootstrap algorithm:

  - Consider the ordered statistics of $T(\mathcal{X}^{* (b)})$, $b = 1, \ldots, B$:
    $$T^*_1 \leq \ldots \leq T^*_B.$$  

  - Let $k_1 = \left[ B \frac{\alpha}{2} \right] + 1$ and $k_2 = \left[ B \left( 1 - \frac{\alpha}{2} \right) \right]$, where $[\cdot]$ stands for the integer part.

  - Finally:
    $$v^*(\frac{\alpha}{2}) = T^*_{k_1}$$
    $$v^*(1 - \frac{\alpha}{2}) = T^*_{k_2}$$

  - NB: Most of the modern softwares have build-in procedures that provide any percentile $v^*(a)$ from the series $T(\mathcal{X}^{* (b)})$, $b = 1, \ldots, B$ (using linear interpolation).

- Finally, if $T(\mathcal{X}) = \theta(F_n)$ (as often the case), the basic bootstrap confidence interval for $\theta(F)$ is given by:

$$\left[ T(\mathcal{X}) + \theta(F_n) - v^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) + \theta(F_n) - v^*(\frac{\alpha}{2}) \right] = \left[ 2 T(\mathcal{X}) - v^*(1 - \frac{\alpha}{2}), 2 T(\mathcal{X}) - v^*(\frac{\alpha}{2}) \right]$$
• **Example:** Let’s come back to the duration data \( n = 10: \)

\[ X = (X_1, \ldots, X_{10}) = (1 5 12 15 20 26 78 145 158 358) \]

We have \( \bar{x} = 81.80 \) and \( s = 112.94. \)

1. **Parametric model:** If \( X \) is exponential (?) we have an exact confidence interval and a pivotal root (available in this simple case).

\[
\frac{\bar{X}}{\mu} \sim \text{Gamma} \left( n, \frac{1}{n} \right)
\]

So that:

\[
\mu \in \left[ \frac{\bar{x}}{q(1 - \frac{\alpha}{2})}, \frac{\bar{x}}{q(\frac{\alpha}{2})} \right]
\]

where \( q(a) \) is the \( a \)-quantile of a \( \text{Gamma}(n, 1/n) \).

2. **Asymptotic parametric approximation:** If \( X \) is exponential (?) we have an asymptotical pivotal root.

\[
\frac{\bar{X} - \mu}{\mu/\sqrt{n}} \sim \text{AN}(0, 1)
\]

and by Slutsky’s theorem \( \sqrt{n}(\bar{X} - \mu)/\bar{X} \sim \text{AN}(0, 1) \). This provides

\[
\mu \in \bar{x} \pm z_{1-\alpha/2} \frac{\bar{x}}{\sqrt{n}}
\]

3. **Asymptotic Non-parametric approximation:** we have the asymptotical pivotal root \( \sqrt{n}(\bar{X} - \mu)/S \sim \text{AN}(0, 1) \). This provides

\[
\mu \in \bar{x} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}}
\]
4. **Bootstrap approaches:**

(a) **Parametric and Non parametric bootstrap:**

\[
\mu \in \left[ 2\bar{x} - v^*(1 - \frac{\alpha}{2}) , \ 2\bar{x} - v^*(\frac{\alpha}{2}) \right]
\]

where \(v^*(a) = \hat{G}^{-1}(a)\) is the \(a\)-quantile of the empirical distribution of \(\bar{X}^*\)

- for the parametric bootstrap \(X_i^* \sim Expo(\bar{x})\),
- for the nonparametric bootstrap \(X_i^*\) drawn from \(X\).

(b) **Parametric Pivotal-bootstrap:** If \(X\) is exponential (?), we can use the pivotal root \(\bar{X}/\mu\) but use a bootstrap approximation of its density in place of the \(\Gamma(n, \frac{1}{n})\).

\[
\mu \in \left[ \frac{\bar{x}}{q^*(1 - \frac{\alpha}{2})} , \ \frac{\bar{x}}{q^*(\frac{\alpha}{2})} \right]
\]

where \(q^*(a)\) is the \(a\)-quantile of the parametric bootstrap distribution of \(\bar{X}^*/\bar{x}\) where \(X_i^* \sim Expo(\bar{x})\). Note that the latter CI is identical to \([v^*(\alpha/2), v^*(1 - \alpha/2)]\) where \(v^*(\cdot)\) is defined above with \(X_i^* \sim Expo(\bar{x})\).

The results are reported in Table 1.3.
<table>
<thead>
<tr>
<th>method</th>
<th>lower limit</th>
<th>upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric Gamma</td>
<td>47.8788</td>
<td>170.5805</td>
</tr>
<tr>
<td>Parametric Pivotal-bootstrap</td>
<td>47.7969</td>
<td>168.9480</td>
</tr>
<tr>
<td>Parametric Bootstrap</td>
<td>23.6067</td>
<td>123.9947</td>
</tr>
<tr>
<td>Parametric Asymp. Normal</td>
<td>31.1008</td>
<td>132.4992</td>
</tr>
<tr>
<td>Nonparametric Asymp. Normal</td>
<td>11.8014</td>
<td>151.7986</td>
</tr>
<tr>
<td>Nonparametric Bootstrap</td>
<td>6.8000</td>
<td>137.3000</td>
</tr>
</tbody>
</table>

Table 1.3: Confidence intervals, duration data with \( n = 10 \). Here \( B = 5000 \)

- **Example:** Estimation of a variance. We simulate \( n = 10 \) observation from a \( N(50, 3^2) \). We obtain

\[
\mathcal{X} = (48.70 45.00 50.38 50.86 46.56 53.57 53.57 49.89 50.98 50.52)
\]

We have the following statistics \( \bar{x} = 50.0038 \), the unbiased estimate of \( \sigma^2 \) is \( s^2 = (1/(n - 1)) \sum_{i=1}^{n} (x_i - \bar{x})^2 = 7.3456 \) and the plug-in estimate is \( v = (1/n) \sum_{i=1}^{n} (x_i - \bar{x})^2 = 6.6110 \). Figure 1.8 shows the estimated sampling distributions of \( S^2 \), obtained by sampling theory, parametric and nonparametric bootstrap (\( B = 5000 \)).
Figure 1.8: Estimates of the sampling distribution of $S^2$: estimated Gamma (dotted), parametric bootstrap (solid), nonparametric bootstrap (dashed).

- Note that the estimates (nonparametric bootstrap) of the bias are

\[
\text{Bias}^*(S^2) = 0.0575 \quad \text{and} \quad \text{Std}^*(S^2) = 2.7492
\]

\[
\text{Bias}^*(V) = -0.6094 \quad \text{and} \quad \text{Std}^*(V) = 2.4743
\]

So that $v_{corr} = 7.2204$ is very similar to $s^2 = 7.3456$.

The confidence intervals for $\sigma^2$ obtained by the parametric and nonparametric bootstrap distribution of $S^2$ are:

parametric bootstrap = $[-1.7756, 11.7803]$

nonparametric bootstrap = $[1.3987, 12.2273]$
### 1.6 Summary Table of the Bootstrap Ideas

<table>
<thead>
<tr>
<th>REAL WORLD</th>
<th>BOOTSTRAP WORLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F$ unknown</td>
<td>Given $X = (X_1, \ldots, X_n)$, $F_n$ is known</td>
</tr>
<tr>
<td>$\theta(F)$ unknown</td>
<td>$\theta(F_n)$ known</td>
</tr>
<tr>
<td>$\text{DGP} ; F : X \sim F$</td>
<td>$\text{DGP}^* ; F_n : X^* \sim F_n$</td>
</tr>
<tr>
<td>$\mathcal{X} = (X_1, \ldots, X_n)$</td>
<td>$\mathcal{X}^* = (X_1^<em>, \ldots, X_n^</em>)$</td>
</tr>
<tr>
<td>$T(\mathcal{X})$ estimator of $\theta(F)$</td>
<td>$T(\mathcal{X}^*)$ estimator of $\theta(F_n)$</td>
</tr>
<tr>
<td>$G(x) = P_F(T(\mathcal{X}) \leq x)$</td>
<td>$\hat{G}(x) = P_{F_n}(T(\mathcal{X}^*) \leq x)$</td>
</tr>
<tr>
<td>Moments $E_F(T(\mathcal{X}))$</td>
<td>$\text{Moments}$ $E_{F_n}(T(\mathcal{X}^*))$</td>
</tr>
<tr>
<td>$\text{Var}_F(T(\mathcal{X}))$</td>
<td>$\text{Var}_{F_n}(T(\mathcal{X}^*))$</td>
</tr>
<tr>
<td>$W = T(\mathcal{X}) - \theta(F)$</td>
<td>$W^* = T(\mathcal{X}^*) - \theta(F_n)$</td>
</tr>
<tr>
<td>$H(x) = P_F(W \leq x) \Rightarrow u(a) = H^{-1}(a)$</td>
<td>$\hat{H}(x) = P_{F_n}(W^* \leq x) \Rightarrow \hat{u}(a) = \hat{H}^{-1}(a)$</td>
</tr>
<tr>
<td>$\text{CI for } \theta$ $[T(\mathcal{X}) - u(1 - \alpha/2), T(\mathcal{X}) - u(\alpha/2)]$</td>
<td>$\text{Bootstrap CI for } \theta$ $[T(\mathcal{X}) - \hat{u}(1 - \alpha/2), T(\mathcal{X}) - \hat{u}(\alpha/2)]$</td>
</tr>
<tr>
<td>$S = \frac{T(\mathcal{X}) - \theta(F)}{\sigma(\mathcal{X})/\sqrt{n}}$</td>
<td>$S^* = \frac{T(\mathcal{X}^<em>) - \theta(F_n)}{\sigma(\mathcal{X}^</em>)/\sqrt{n}}$</td>
</tr>
<tr>
<td>$K(x) = P_F(S \leq x) \Rightarrow y(a) = K^{-1}(a)$</td>
<td>$\hat{K}(x) = P_{F_n}(S^* \leq x) \Rightarrow \hat{y}(a) = \hat{K}^{-1}(a)$</td>
</tr>
<tr>
<td>$\text{CI for } \theta$ $[T(\mathcal{X}) - y(1 - \alpha/2)\frac{\sigma(\mathcal{X})}{\sqrt{n}}, T(\mathcal{X}) - y(\alpha/2)\frac{\sigma(\mathcal{X})}{\sqrt{n}}]$</td>
<td>$\text{Bootstrap CI for } \theta$ $[T(\mathcal{X}) - \hat{y}(1 - \alpha/2)\frac{\sigma(\mathcal{X}^<em>)}{\sqrt{n}}, T(\mathcal{X}) - \hat{y}(\alpha/2)\frac{\sigma(\mathcal{X}^</em>)}{\sqrt{n}}]$</td>
</tr>
</tbody>
</table>
Chapter 2

More on Confidence Intervals

2.1 The Percentile Method

Suppose there exist some unknown monotone transformation of 
\( T(\mathcal{X}) = \theta(F_n) \), say \( U = h(T) \) which has a symmetric distribution 
around \( \eta = h(\theta) \). So, \( h(T) \) is an unbiased estimator of \( h(\theta) \).

- If we knew \( h \), a confidence interval could be obtained for \( \eta \) as 
  \[
  \left[ U(\mathcal{X}) - u_U(1 - \alpha/2), U(\mathcal{X}) - u_U(\alpha/2) \right],
  \]
  where \( u_U(a) \) is the \( a \)-quantile of the sampling distribution of 
  \( U - \eta \).

- Due to the symmetry of the sampling distribution of \( U - \eta \) 
  around 0, we have here 
  \[
  -u_U(\alpha/2) = u_U(1 - \alpha/2)
  \]
  so the confidence interval can be written as 
  \[
  \left[ U(\mathcal{X}) + u_U(\alpha/2), U(\mathcal{X}) + u_U(1 - \alpha/2) \right]
  \]
• Basic Bootstrap approximation:

– If we knew \( h \), we could use the basic bootstrap method to obtain an approximation of the confidence interval for \( \eta \) as follows:

\[
[U(\mathcal{X}) + u^*_U(\alpha/2), U(\mathcal{X}) + u^*_U(1 - \alpha/2)]
\]

where \( u^*_U(a) \) is the \( a \)-quantile of the bootstrap distribution of \( U^* - \hat{\eta} \), with \( U^* = U(\mathcal{X}^*) = h(T(\mathcal{X}^*)) \) and \( \hat{\eta} = U(\mathcal{X}) = h(T(\mathcal{X})) \).

– Since \( u^*_U(a) = v^*_U(a) - U(\mathcal{X}) \) this is equivalent to

\[
\eta \in \left[ v^*_U(\frac{\alpha}{2}), v^*_U(1 - \frac{\alpha}{2}) \right]
\]

where \( v^*_U(a) \) is the \( a \)-quantile of the bootstrap distribution of \( U(\mathcal{X}^*) \).

• Due to the monotonicity of \( h \), this confidence interval for \( \eta \) can be transformed back to a confidence interval for \( \theta \),

– We have

\[
v^*_U(a) = h(v^*(a)),
\]

where \( v^*(a) \) is the \( a \)-quantile of the bootstrap distribution of \( T(\mathcal{X}^*) \).

– We obtain for \( \theta(F) \) the bootstrap percentile interval:

\[
\theta(F) \in \left[ v^*(\frac{\alpha}{2}), v^*(1 - \frac{\alpha}{2}) \right]
\]
• This does not depend on $h$ and so can be implemented without knowing $h$. Note that the bootstrap percentile interval has the **transformation invariance property** (for any monotone increasing mapping).

• It is also called a **backward** bootstrap interval since it looks for critical values in the “wrong” tails. Indeed, compare in the $\theta$ scale:

  Percentile $\left[ T(\mathcal{X}) + u^*(\frac{\alpha}{2}), T(\mathcal{X}) + u^*(1 - \frac{\alpha}{2}) \right]$

  Basic bootstrap $\left[ T(\mathcal{X}) - u^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) - u^*(\frac{\alpha}{2}) \right]$

• It is more popular than the basic bootstrap confidence interval because it works very well in many applications. It is more stable to original sampling variations. Compare both intervals:

  Percentile $\left[ v^*(\frac{\alpha}{2}), v^*(1 - \frac{\alpha}{2}) \right]$

  Basic bootstrap $\left[ 2T(\mathcal{X}) - v^*(1 - \frac{\alpha}{2}), 2T(\mathcal{X}) - v^*(\frac{\alpha}{2}) \right]$.

• **Remarks:**

  – The percentile interval is always contained in the domain of $\theta(F)$, because $v^*(a)$ is in the domain of $T(\mathcal{X})$.

  – The percentile interval is correct if and only if such a monotone transformation $h$ exists (but we do not need to know it).

  – The symmetry of the distribution in the transformed scale implies that $U$ is an **unbiased** estimator of $\eta$.

  – If $h(T) = T$, Basic bootstrap $\approx$ Percentile.
2.2 The BC (bias-corrected) Percentile Method

• Suppose there exist some unknown monotone (increasing) transformation of $T(\mathcal{X}) = \theta(F_n)$, say $U = h(T)$ which has (at least asymptotically) a Normal distribution around $\eta - w$ where $\eta = h(\theta)$. In a sense $w$ is the “bias” of $U(\mathcal{X})$:

$$h(T) \sim N(h(\theta) - w, 1)$$
$$U - \eta \sim N(-w, 1)$$

Note that the scaling factor to obtain a variance equal to 1 may be introduced in the monotone transformation $h$ (if $\text{Var}(U)$ does not depend on $\eta$: otherwise we should use the $BC_a$ method below).

• How to estimate $w$?

  – Since $U - \eta + w = Z \sim N(0, 1)$ we have:

    $$P(U > \eta) = P(Z - w > 0) = P(Z > w).$$

  – Now, due to the (increasing) monotonicity of $h$:

    $$P(U > \eta) = P(h(T) > h(\theta)) = P(T > \theta) = P(Z > w)$$

  – This involve an equation in $w$:

    $$P(T \leq \theta) = P(Z \leq w) = \Phi(w)$$

    the solution is:

    $$w = \Phi^{-1}(P(T \leq \theta)).$$
– The bootstrap estimate of $w$ is then:

\[ \hat{w} = \Phi^{-1}(P(T(X^*) \leq \theta(F_n) \mid X)) \]

\[ = \Phi^{-1}(P_{F_n}(T(X^*) \leq \theta(F_n))) \]

\[ = \Phi^{-1}(\hat{G}(\theta(F_n))) \]

\[ = \Phi^{-1}\left(\frac{\# \left[ T(X^{*\text{(b)}}) \leq \theta(F_n) \right]}{B}\right) \]

• How to obtain confidence interval for $\theta$?

– Since $U - \eta + w = Z \sim N(0, 1)$, we have:

\[ P\left(z_{\alpha/2} \leq U - \eta + w \leq z_{1-\alpha/2}\right) = 1 - \alpha \]

\[ P\left(U + w - z_{1-\alpha/2} \leq \eta \leq U + w - z_{\alpha/2}\right) = 1 - \alpha \]

\[ P\left(U + w + z_{\alpha/2} \leq \eta \leq U + w + z_{1-\alpha/2}\right) = 1 - \alpha \]

\[ P\left(\hat{\eta}_{\alpha/2} \leq \eta \leq \hat{\eta}_{1-\alpha/2}\right) = 1 - \alpha, \]

where $\hat{\eta}_a = U + w + z_a$.

– Finally, confidence interval for $\theta$: use the inverse transform.

\[ P\left(\hat{\theta}_{\alpha/2} \leq \theta \leq \hat{\theta}_{1-\alpha/2}\right) = 1 - \alpha \]

where $\hat{\theta}_a = h^{-1}(\hat{\eta}_a)$.

• But $h(\cdot)$ is unknown, so is $\hat{\theta}_a$. 
• We will use the Normality of $U$, to evaluate $\hat{\theta}_a$.

  - We have

    \[
    \hat{G}(\hat{\theta}_a) = P \left( T(\mathcal{X}^*) \leq \hat{\theta}_a | \mathcal{X} \right) = P \left( U(\mathcal{X}^*) \leq \hat{\eta}_a | \mathcal{X} \right)
    \]

  - Since $\hat{\eta}_a = U(\mathcal{X}) + w + z_a$ this gives:

    \[
    \hat{G}(\hat{\theta}_a) = P \left( U(\mathcal{X}^*) - U(\mathcal{X}) + w \leq 2w + z_a | \mathcal{X} \right), \text{ Boot. world}
    \approx P_F \left( U(\mathcal{X}) - \eta + w \leq 2w + z_a \right), \text{ Real world}
    \]

  - Now, since $U - \eta + w = Z \sim N(0,1)$, we have:

    \[
    \hat{G}(\hat{\theta}_a) \approx \Phi (2w + z_a)
    \]

• Finally we obtain

  \[
  \hat{\theta}_a = \hat{G}^{-1} \left( \Phi (2w + z_a) \right)
  = v^* (\Phi (2w + z_a))
  \]

  where $w$ will be estimated by $\hat{w}$:

  \[
  \hat{\theta}_a = v^* \left( \Phi \left[ 2 \Phi^{-1} \left( \frac{\#[T(\mathcal{X}^*^{(b)}) \leq \theta(F_n)]}{B} \right) + z_a \right] \right)
  \]

• Note that if $w = 0$, the BC method coincides with the usual percentile method: $\hat{\theta}_a = v^*(a)$. 
2.3 The $BC_a$ Percentile Method

- The basic idea

Suppose there exist some unknown monotone (increasing) transformation $h(\cdot)$ of $\hat{\theta} = T(\mathcal{X})$, say $\hat{\eta} = h(T(\mathcal{X}))$ which has (at least asymptotically) a Normal distribution around $\eta = h(\theta)$ with a bias $w$ and a variance which may depend on $\eta$:

$$H_{BC_a} : \frac{\hat{\eta} - \eta}{\sigma_{\eta}} \sim N(-w, 1),$$

where $\sigma_{\eta} = \sigma_{\eta_0}[1+a(\eta-\eta_0)]$ where $\eta_0$ is any convenient reference point on the scale of $\eta$ values: the idea is that we approximate (when $n$ is large) $\sigma_{\eta}$ by a linear function. To simplify we will chose below $\eta_0$ such that $\sigma_{\eta_0} = 1$. Again, the scaling factor to obtain the variance 1 in the normal approximation can be tuned in the monotone function $h$.

As above ($BC$-percentile method) $w$ accounts for a possible bias in $\hat{\eta}$ as estimator of $\eta$ but we allow here the variance of $\hat{\eta}$ being a function of $\eta$: here $a$ is called the acceleration factor (rate of change of the std deviation of $\hat{\eta}$ as a function of $\eta$).

The hypothesis $H_{BC_a}$ holds for a large class of problems where the error in the approximation is of order $O_p(n^{-1})$ (usually, CLT holds with an error of order $O_p(n^{-1/2})$).
• Confidence intervals for $\eta$

- From $H_{BC_\alpha}$, we obtain for large $n$:

$$\text{Prob}(\hat{\eta} - \eta + w\sigma_\eta \leq z_{1-\alpha/2}\sigma_\eta) = 1 - \alpha/2.$$ 

Since $\sigma_\eta - \sigma_{\hat{\eta}} = a(\eta - \hat{\eta})$ we have:

$$\text{Prob}(\hat{\eta} - \eta + w\sigma_{\hat{\eta}} + aw(\eta - \hat{\eta}) \leq z_{1-\alpha/2}(\sigma_{\hat{\eta}} + a(\eta - \hat{\eta})) = 1 - \alpha/2,$$

from which it is easy to derive (by noticing that $z_{1-\alpha/2} = -z_{\alpha/2}$):

$$\text{Prob} \left( \eta \leq \hat{\eta} + \sigma_{\hat{\eta}} \frac{w + z_{\alpha/2}}{1 - a(w + z_{\alpha/2})} \right) = \alpha/2.$$ 

- A $(1 - \alpha) \times 100\%$ level confidence interval for $\eta$ is thus obtained by

$$\text{Prob}(\hat{\eta}_{\alpha/2} \leq \eta \leq \hat{\eta}_{1-\alpha/2}) = 1 - \alpha,$$

where $\hat{\eta}_{\beta} = \hat{\eta} + \sigma_{\hat{\eta}} \frac{w + z_{\beta}}{1 - a(w + z_{\beta})}$.

• Confidence intervals for $\theta$

We have to come back to the $\theta$ scale.

- Since $h$ is monotone we have

$$\text{Prob}(\hat{\theta}_{\alpha/2} \leq \theta \leq \hat{\theta}_{1-\alpha/2}) = 1 - \alpha,$$

where $\hat{\theta}_{\beta} = h^{-1}(\hat{\eta}_{\beta})$. But $h$ and $h^{-1}$ are unknown.
Let \( \hat{G}(\cdot) \) be the cdf of \( \hat{\theta}^* = T(\mathcal{X}^*) \). We have, in the bootstrap world:

\[
\hat{G}(\hat{\theta}_\beta) = \text{Prob}(\hat{\theta}^* \leq \hat{\theta}_\beta | \mathcal{X}) = \text{Prob}(\hat{\eta}^* \leq \hat{\eta}_\beta | \mathcal{X})
\]

\[
= \text{Prob}\left( \frac{\hat{\eta}^* - \hat{\eta}}{\hat{\sigma}_\hat{\eta}} + w \leq w + \frac{w + z_\beta}{1 - a(w + z_\beta)} \right)
\]

This is the bootstrap analog of the real world quantity:

\[
\text{Prob}_F\left( \frac{\hat{\eta}^* - \eta}{\sigma_\eta} + w \leq w + \frac{w + z_\beta}{1 - a(w + z_\beta)} \right),
\]

so (by \( H_{BC_a} \)) we obtain:

\[
\hat{G}(\hat{\theta}_\beta) \approx \Phi\left( w + \frac{w + z_\beta}{1 - a(w + z_\beta)} \right),
\]

from which we derive an estimate of \( \hat{\theta}_\beta \):

\[
\hat{\theta}_\beta = \hat{G}^{-1}\left[ \Phi\left( w + \frac{w + z_\beta}{1 - a(w + z_\beta)} \right) \right]
\]

\[
= v^*\left[ \Phi\left( w + \frac{w + z_\beta}{1 - a(w + z_\beta)} \right) \right],
\]

where \( v^*(\cdot) \) is the quantile of the bootstrap distribution of \( T(\mathcal{X}^*) \).

We remark that if \( a = 0 \) we are back to the \( BC \)-percentile method and if \( a = w = 0 \) we are back to the percentile method. The 3 methods are looking to the same bootstrap distribution of \( T(\mathcal{X}^*) \), but the levels of the quantiles are adjusted to correct for bias (\( BC \)) or to correct for bias and acceleration (\( BC_a \)).
• Estimation of $w$ and $a$

  – For the bias $w$, the same as the $BC$ method:

    \[
    G(\theta) = \text{Prob}_F(\hat{\theta} \leq \theta) = \text{Prob}_F(\hat{\eta} \leq \eta)
    \]
    \[
    = \text{Prob}_F(\frac{\hat{\eta} - \eta}{\sigma_{\eta}} \leq 0) = \text{Prob}(Z \leq w) = \Phi(w)
    \]

    So that $w = \Phi^{-1}(G(\theta))$ which can be estimated by the bootstrap approximation:

    \[
    \hat{w} = \Phi^{-1}(\text{Prob}(\hat{\theta}^* \leq \hat{\theta} | \mathcal{X})) = \Phi^{-1}(\hat{G}(\hat{\theta}))
    \]
    \[
    = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^* \leq \hat{\theta}\}}{B}\right).
    \]

  – For the acceleration $a$, the method is very easy to use but its justification is more elaborated (see Efron-Tibshirani, Chapter 22). A consistent nonparametric estimator of $a$ is given by:

    \[
    \hat{a} = \frac{\sum_{i=1}^{n} L^3_{\theta}(x_i, F_n)}{6\left[\sum_{i=1}^{n} L^2_{\theta}(x_i, F_n)\right]^{3/2}},
    \]

    where $L_{\theta}(x_i, F_n)$ is the empirical influence component defined in Section 1.3. An easy way to estimate it (see Chapter 7 below) is by using the $i$th jacknife value:

    \[
    \hat{a} = \frac{\sum_{i=1}^{n} (\hat{\theta}(\cdot) - \hat{\theta}(i))^3}{6\left[\sum_{i=1}^{n} (\hat{\theta}(\cdot) - \hat{\theta}(i))^2\right]^{3/2}},
    \]

    where $\hat{\theta}(i) = T(\mathcal{X}(i))$ is the jacknife value of $\hat{\theta}$ obtained by applying the statistics $T(\cdot)$ to the original sample with the $i$th point $X_i$ deleted (leave-one-out), and $\hat{\theta}(\cdot) = \sum_{i=1}^{n} \hat{\theta}(i)/n$ is their average.


2.4 The Bootstrap-\(t\) Method

2.4.1 Studentization in the Real World

• Suppose we have access to (asymptotic) pivotal roots (MLE, or by Delta method, \ldots):

\[
S = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\sigma(F)} \right)
\]

where \(\sigma^2(F) = n \text{Var}_F(T(\mathcal{X}))\).

– Suppose we know the (asymptotic) sampling distribution of \(S\)

\[
H(x) = P(S \leq x) = P \left\{ \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\sigma(F)} \right) \leq x \right\}.
\]

Let \(x_a\) be the \(a\)-quantile of \(H\): \(x_a = H^{-1}(a)\).

– If \(\sigma\) is known: confidence interval for \(\theta(F)\):

\[
\theta(F) \in \left[ T(\mathcal{X}) - n^{-1/2} \sigma x_{1-a/2}, T(\mathcal{X}) - n^{-1/2} \sigma x_{\alpha/2} \right]
\]

• If \(\sigma\) is unknown: “Studentization”

– We often have a consistent estimator \(\hat{\sigma}(\mathcal{X})\). Then the pivotal root is:

\[
U = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\hat{\sigma}(\mathcal{X})} \right)
\]

– If we know the (asymptotic) sampling distribution of \(U\)

\[
K(x) = P(U \leq x) = P \left\{ \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\hat{\sigma}(\mathcal{X})} \right) \leq x \right\}.
\]
– The studentized confidence interval for $\theta(F)$ is:

$$
\theta(F) \in \left[ T(\mathcal{X}) - n^{-1/2} \hat{\sigma}(\mathcal{X}) y_{1-\alpha/2}, T(\mathcal{X}) - n^{-1/2} \hat{\sigma}(\mathcal{X}) y_{\alpha/2} \right]
$$

where $y_a$ be the $a$-quantile of $K$: $y_a = K^{-1}(a)$.

– Note that if asymptotic is used, $K \equiv H$ (Slutsky’s theorem)

2.4.2 Studentization in the Bootstrap World

• Bootstrap approximations:

– Bootstrap estimates of $x_a$ provided by the $a$-quantiles of $\hat{H}$:

$$
\hat{H}(x) = P(S^* \leq x \mid \mathcal{X}) = P\left( \sqrt{n} \left( \frac{T(\mathcal{X}^*) - \theta(F_n))}{\sigma(F_n)} \right) \leq x \mid \mathcal{X} \right).
$$

– Bootstrap estimates of $y_a$ provided by the $a$-quantiles of $\hat{K}$:

$$
\hat{K}(x) = P(U^* \leq x \mid \mathcal{X}) = P\left( \sqrt{n} \left( \frac{T(\mathcal{X}^*) - \theta(F_n))}{\hat{\sigma}(\mathcal{X}^*)} \right) \leq x \mid \mathcal{X} \right).
$$

• Bootstrap confidence intervals:

– If $\sigma$ is known:

$$
\theta(F') \in \left[ T(\mathcal{X}) - n^{-1/2} \hat{\sigma} \hat{x}_{1-\alpha/2}, T(\mathcal{X}) - n^{-1/2} \hat{\sigma} \hat{x}_{\alpha/2} \right],
$$

where $\hat{x}(a)$ is the $a$-quantile of $\hat{H}(x)$. 

- If $\sigma$ unknown: usual case (Bootstrap-$t$ CI)

$$
\theta(F) \in \left[T(\mathcal{X}) - n^{-1/2}\hat{\sigma}(\mathcal{X})\hat{y}_{1-\alpha/2}, T(\mathcal{X}) - n^{-1/2}\hat{\sigma}(\mathcal{X})\hat{y}_{\alpha/2}\right]
$$

where $\hat{y}(a)$ is the $a$-quantile of $\hat{K}(x)$.

- **Accuracy of the CI obtained by bootstrap methods**

  - An approximate end point of level $\alpha$, say $\hat{\theta}_\alpha$ is such that $\text{Prob}(\theta < \hat{\theta}_\alpha) \approx \alpha$.

    - It is said as being **First-order accurate** if
      $$
      \text{Prob}(\theta < \hat{\theta}_\alpha) = \alpha + O(n^{-1/2}).
      $$

    - It is said as being **Second-order accurate** if
      $$
      \text{Prob}(\theta < \hat{\theta}_\alpha) = \alpha + O(n^{-1}).
      $$

  - In summary and in general (under regularity conditions: “when the bootstrap works”):

    - Large sample (Normal approximation) is first order accurate
    - Basic Bootstrap CIs are first order accurate
    - Percentile (when appropriate) is first order accurate
    - Bootstrap-$t$ (when available) is second order accurate
    - $BC_a$ percentile (when appropriate) is second order accurate
    - Double bootstrap (see later: estimation of the variance by an inner bootstrap loop, allowing the use of the bootstrap-$t$) are second order accurate

  - More will be said in Chapters 3 and 6.
• **Example:** Let’s come back to the duration data $n = 10$:

$$\mathcal{X} = (X_1, \ldots, X_{10}) = (1, 5, 12, 15, 20, 26, 78, 145, 158, 358)$$

We have $\bar{x} = 81.80$ and $s = 112.94$.

- Here we have:

$$U = \sqrt{n} \left( \frac{\bar{X} - \mu}{s} \right) \sim AN(0, 1)$$

The bootstrap-$t$ confidence interval is given by

$$\mu \in \left[ \bar{X} - n^{-1/2} s \hat{y}_{1-\alpha/2}, \bar{X} - n^{-1/2} s \hat{y}_{\alpha/2} \right],$$

where $\hat{y}(a)$ is the $a$-quantile of $\hat{K}(x)$, the empirical bootstrap distribution of

$$U^* = \sqrt{n} \left( \frac{\bar{X}^* - \bar{X}}{S^*} \right)$$

- The results of Table 1.3 are now completed by the percentile, $BC$-percentile ($\hat{w} = 0.0843$), $BC_a$-percentile ($\hat{a} = 0.0846$) and the nonparametric bootstrap-$t$.

<table>
<thead>
<tr>
<th>method</th>
<th>lower limit</th>
<th>upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric Gamma</td>
<td>47.8788</td>
<td>170.5805</td>
</tr>
<tr>
<td>Parametric Pivotal-bootstrap</td>
<td>47.7969</td>
<td>168.9480</td>
</tr>
<tr>
<td>Parametric Bootstrap</td>
<td>23.6067</td>
<td>123.9947</td>
</tr>
<tr>
<td>Parametric Asymp. Normal</td>
<td>31.1008</td>
<td>132.4992</td>
</tr>
<tr>
<td>Nonparametric Asymp. Normal</td>
<td>11.8014</td>
<td>151.7986</td>
</tr>
<tr>
<td>Nonparametric Basic Bootstrap</td>
<td>6.8000</td>
<td>137.3000</td>
</tr>
<tr>
<td>Nonparametric Percentile</td>
<td>26.2500</td>
<td>156.8000</td>
</tr>
<tr>
<td>Nonparametric $BC$-Percentile</td>
<td>28.7000</td>
<td>164.2567</td>
</tr>
<tr>
<td>Nonparametric $BC_a$-Percentile</td>
<td>35.3754</td>
<td>184.9024</td>
</tr>
<tr>
<td>Nonparametric Bootstrap-$t$</td>
<td>23.3073</td>
<td>230.1239</td>
</tr>
</tbody>
</table>

Table 2.1: *Confidence intervals, duration data with $n = 10$. Here $B = 5000*
2.5 Some Simulated Examples

- **Example 1**: $X_i \sim \text{Expo} (\mu)$, simulated data with $\mu = 10$.

  - One simulated sample, $n = 10$, with $\bar{X} = 7.7717$ and $S^2 = (4.3045)^2$. Comparison of CI for $\mu$ at the level 95%. For the bootstrap $B = 5000$. Results in Table 2.2.

<table>
<thead>
<tr>
<th>method</th>
<th>lower limit</th>
<th>upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated Gamma</td>
<td>4.5489</td>
<td>16.2066</td>
</tr>
<tr>
<td>Estimated As. Normal</td>
<td>2.9548</td>
<td>12.5885</td>
</tr>
<tr>
<td>Basic Bootstrap</td>
<td>5.1181</td>
<td>10.1953</td>
</tr>
<tr>
<td>Percentile Bootstrap</td>
<td>5.3481</td>
<td>10.4253</td>
</tr>
</tbody>
</table>

Table 2.2: *Comparisons of confidence intervals, exponential case with $n = 10$ and $\mu = 10$.*

- **Precision in Monte-Carlo experiments:**
  - How is the precision for the estimation of the coverage probabilities when using a Monte-Carlo experiment?
  - If we estimate a proportion $p$, over a sample of size $M$, the Monte-Carlo precision (at level 95%) is roughly
    \[
    \text{Error bound} = 2 \sqrt{\frac{p(1-p)}{M}}.
    \]
  - If we estimate $p \approx 0.95$ we have:
    \[
    \text{Error bound} = 2 \sqrt{\frac{0.05 \times 0.95}{1000}} = 0.0138
    \]
  - Table 2.3 displays some values of the error bound as a function of $M$. 

Table 2.3: Error bound for the Monte-Carlo precision when estimating a probability $p = 0.95$.

- Evaluation of the performance of confidence intervals by Monte-Carlo experiments:
  - we repeat $M = 1000$ times the simulation of a random sample of size $n$ as described above for one sample
  - we evaluate the average length of the intervals and the coverage probabilities (proportion of time, under the 1000 replication, that the CI contains the true $\mu = 10$).
  - See Table 2.4.
  - The MC-precision (error bound) for the coverages is thus $0.0138$.
  - The asymptotic Normal performs well: CLT for $\bar{X}$.

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 5$</th>
<th>$n = 10$</th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
</tr>
<tr>
<td>Exact Gamma</td>
<td>26.25</td>
<td>0.944</td>
<td>14.90</td>
<td>0.958</td>
</tr>
<tr>
<td>As. Normal</td>
<td>17.76</td>
<td>0.874</td>
<td>12.32</td>
<td>0.908</td>
</tr>
<tr>
<td>Basic Boot.</td>
<td>13.39</td>
<td>0.765</td>
<td>10.68</td>
<td>0.839</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>13.39</td>
<td>0.782</td>
<td>10.68</td>
<td>0.865</td>
</tr>
</tbody>
</table>

Table 2.4: Performances of confidence intervals for $\mu = 10$: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 5000$. 
• **Example 2:** Same Exponential Model $X \sim \text{Expo} (\mu)$: Estimation of the probability of “no event” in one unit time period. We have an iid sample $(X_1, \ldots, X_n)$ of exponential duration times.

- Let $N$ be the number of “event” in one unit time period, $N \sim \text{Poisson}(1/\mu)$

  - Parameter of interest
    
    $$
    \pi = P(N = 0) = \exp(-1/\mu)
    $$

  - The MLE of $\pi$ is $\hat{\pi} = \exp(-1/\bar{X})$.

  - What are the sampling properties of $\hat{\pi}$? We have here (delta method) an asymptotic result:
    
    $$
    \hat{\pi} \sim \text{AN} \left( \pi, \frac{\pi^2}{n \mu^2} = \frac{1}{n} \left( \frac{\pi}{\mu} \right)^2 \right)
    $$

    We see $\sigma(\hat{\pi}) = \frac{e^{-1/\mu}}{\sqrt{n \mu}}$, it will be estimated by $\hat{\sigma}(\hat{\pi}) = \frac{e^{-1/\bar{x}}}{\sqrt{n \bar{x}}}$.

- Figure 2.1 show the bootstrap sampling distribution of $\hat{\pi}$ and the estimated asymptotic normal (here $\pi = 0.9048$) for one sample with $n = 10$, $\hat{\pi}_{obs} = 0.9161$ and $n = 100$, $\hat{\pi}_{obs} = 0.9136$. 

- Monte-Carlo evaluation of the performances of confidence intervals for $\pi$.

- The asymptotic CI for $\pi$ of level $(1 - \alpha)$ is:

$$
\left[ \hat{\pi} - z_{1-\alpha/2} \frac{e^{-1/\bar{x}}}{\sqrt{n\bar{x}}}, \hat{\pi} + z_{1-\alpha/2} \frac{e^{-1/\bar{x}}}{\sqrt{n\bar{x}}} \right]
$$

- For the bootstrap-t we have

$$
\left[ \hat{\pi} - \hat{y}_{1-\alpha/2} \frac{e^{-1/\bar{x}}}{\sqrt{n\bar{x}}}, \hat{\pi} - \hat{y}_{\alpha/2} \frac{e^{-1/\bar{x}}}{\sqrt{n\bar{x}}} \right]
$$

where $\hat{y}_a$ is the $a$-quantile of

$$
\hat{K}(x) = P \left( \sqrt{n} \left( \frac{\hat{\pi}^* - \hat{\pi}}{\hat{\pi}^*/X^*} \right) \leq x \mid \mathcal{X} \right).
$$

- The results are in Table 2.5.
<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 5$</th>
<th>$n = 10$</th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
</tr>
<tr>
<td>As. Normal</td>
<td>0.186</td>
<td>0.956</td>
<td>0.122</td>
<td>0.957</td>
</tr>
<tr>
<td>Basic Boot.</td>
<td>0.217</td>
<td>0.817</td>
<td>0.130</td>
<td>0.879</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>0.217</td>
<td>0.801</td>
<td>0.130</td>
<td>0.874</td>
</tr>
<tr>
<td>BC-perc.</td>
<td>0.198</td>
<td>0.808</td>
<td>0.124</td>
<td>0.883</td>
</tr>
<tr>
<td>Bootstrap-t</td>
<td>0.155</td>
<td>0.857</td>
<td>0.107</td>
<td>0.908</td>
</tr>
</tbody>
</table>

Table 2.5: Performances of confidence intervals for $\pi = 0.9048$: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 2000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.

- **Example 3:** Correlation coefficient.

\[ X \sim N_2 \left( \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 & 1.5 \\ 1.5 & 1 \end{pmatrix} \right) \]

Here $\rho = 0.75$. The MLE estimator is $\hat{\rho} = r$.

- No simple results for the sampling distribution of $\hat{\rho}$ (see Delta method above)

- We also have in this case a simple asymptotical pivotal root (Fisher z-transfrom):

\[ \hat{z} = \tanh^{-1}(r) \sim AN(\tanh^{-1}(\rho), \frac{1}{n-3}) \]

where $\tanh^{-1}(x) = \frac{1}{2} \log \left( \frac{1+x}{1-x} \right)$ and $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.

- For one sample with $n = 100$, where $r = 0.7644$ see Figure 2.2
Figure 2.2: Bootstrap approximation for the sampling distribution of $\hat{\rho} = r$ (left) and of $\hat{z} = \tanh^{-1}(r)$ (right), original sample has $n = 100$ and $r = 0.7644$ whereas the true $\rho = 0.75$. Dot is the true asymptotic normal (with $\rho = 0.75$), dashed is the estimated Normal approximation (with $r = 0.7644$) and solid is the Bootstrap approximation. Here $B = 2000$.

- Monte-Carlo evaluation of the performances of CI for $\rho$. The asymptotic CI for $\rho$ of level $(1 - \alpha)$ is:

$$\rho \in \left[ \tanh(\hat{z} - z_{1-\alpha/2}/\sqrt{n - 3}), \tanh(\hat{z} + z_{1-\alpha/2}/\sqrt{n - 3}) \right]$$

- Results in Table 2.6.

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 10$</th>
<th>$n = 20$</th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
</tr>
<tr>
<td>As. Normal</td>
<td>0.68</td>
<td>0.953</td>
<td>0.43</td>
</tr>
<tr>
<td>Basic Boot.</td>
<td>0.65</td>
<td>0.778</td>
<td>0.40</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>0.65</td>
<td>0.922</td>
<td>0.40</td>
</tr>
<tr>
<td>BC-Perc.</td>
<td>0.73</td>
<td>0.944</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 2.6: Performances of confidence intervals for $\rho = 0.75$: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$. 


• **Example 4:** An other example for correlation

- Scenario:
  - Let $X_1 \sim N(0, 1)$ and $X_2 \equiv X_1^2$

- Covariance:
  \[
  Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) \\
  = E(X_1^3) - 0 = 0
  \]

- $\rho = 0$ but $X_1$ and $X_2$ are not independent!

- Monte-Carlo evaluation of different confidence intervals see Table 2.7

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 10$</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
</tr>
<tr>
<td>As. Normal</td>
<td>0.95</td>
<td>0.628</td>
<td>0.51</td>
<td>0.616</td>
</tr>
<tr>
<td>Basic Boot.</td>
<td>1.49</td>
<td>0.431</td>
<td>0.95</td>
<td>0.722</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>1.49</td>
<td>0.847</td>
<td>0.95</td>
<td>0.891</td>
</tr>
<tr>
<td>BC-Perc.</td>
<td>1.58</td>
<td>0.893</td>
<td>0.95</td>
<td>0.898</td>
</tr>
</tbody>
</table>

Table 2.7: Performances of confidence intervals for $\rho = 0$: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$. 
– For one sample with $n = 100$, where $r = -0.3856$ see Figure 2.3

Figure 2.3: Bootstrap approximation for the sampling distribution of $\hat{\rho} = r$ (left) and of $\hat{z} = \tanh^{-1}(r)$ (right), original sample has $n = 100$ and $r = 0.0942$ whereas the true $\rho = 0$. Dot is the “true” but wrong asymptotic normal (with $\rho = 0$), dashed is the estimated Normal approximation (with $r = -0.3856$) and solid is the Bootstrap approximation. Here $B = 2000$.

– For one sample with $n = 500$, where $r = -0.0416$ see Figure 2.4

Figure 2.4: Bootstrap approximation for the sampling distribution of $\hat{\rho} = r$ (left) and of $\hat{z} = \tanh^{-1}(r)$ (right), original sample has $n = 500$ and $r = -0.1183$ whereas the true $\rho = 0$. Ddot is the “true” but wrong asymptotic normal (with $\rho = 0$), dashed is the estimated Normal approximation (with $r = -0.0416$) and solid is the Bootstrap approximation. Here $B = 2000$. 
• Example 5: Failure of the Bootstrap.

\[ X_i \sim U(0, \theta) \]

- The problem is to estimate \( \theta \).
- MLE \( \hat{\theta} \):

\[ \hat{\theta} = X_{(n)} = \max(X_1, \ldots, X_n) \]

- Here we know the exact sampling distribution of \( \hat{\theta} \)

\[ \hat{\theta} = X_{(n)} = \max(X_1, \ldots, X_n) \]

- Let’s try the nonparametric bootstrap, with one simulated sample of size \( n \) with a true value of \( \theta = 1 \)

- Here \( f(x_{(n)}) = nx_{(n)}^{n-1} \) on \([0, 1]\)
- Results in Figure 2.5 with \( n = 100 \) and \( n = 500 \).

Figure 2.5: Bootstrap approximation for the sampling distribution of \( X_{(n)} \), original sample has \( n = 100 \) and \( x_{(n)} = 0.9806 \) (left) and \( n = 500 \) and \( x_{(n)} = 0.9966 \) (right). Here \( B = 2000 \). Solid line is the bootstrap approximation, dashdot line is the true sampling density.

The bootstrap does not work here !!
- The problem: the proportion of $X_{(n)}^*(b)$ equal to $x_{(n)}$ over the $B = 2000$ replications is equal to 0.6285 in Figure 2.6 (left) and to 0.6520 in Figure 2.6 (right): see next chapter for more details.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2_6.pdf}
\caption{Figure 2.6:}
\end{figure}

\textbf{But there are solutions:} smoothed bootstrap and subsampling.

- \textit{Smoothed} bootstrap: resample from a nonparametric density estimator of $f(x)$. See Figure 2.7: We see indeed that the distribution has recovered the shape, but there is a shift due to the inherent bias of the estimator.

- Left: original sample has $n = 100$ and $x_{(n)} = 0.9806$. $Bias^* = -0.0058$ and $Std^* = 0.0055$

- Right: original sample has $n = 500$ and $x_{(n)} = 0.9966$. $Bias^* = -0.0024$ and $Std^* = 0.0023$

- As shown in Table 2.8, this distribution allows to recover the coverages for the basic bootstrap
- The percentile approaches (even done on bias corrected estimators) do not work: we are far from a symmetric distribution for $X_{(n)}$.

![Graph](image)

Figure 2.7: Smoothed-Bootstrap approximation of the sampling distribution of $X_{(n)}$, original sample has $n = 100$ and $x_{(n)} = 0.9806$ (left) and $n = 500$ and $x_{(n)} = 0.9966$ (right). Here $B = 2000$. Solid line is the bootstrap approximation, dashdot line is the true sampling density.

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 10$</th>
<th>$n = 50$</th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
</tr>
<tr>
<td>Basic Boot.</td>
<td>0.2865</td>
<td>0.8790</td>
<td>0.0748</td>
<td>0.9340</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>0.2865</td>
<td>0.5480</td>
<td>0.0748</td>
<td>0.6140</td>
</tr>
<tr>
<td>BC-Perc.</td>
<td>0.3895</td>
<td>0.5150</td>
<td>0.1077</td>
<td>0.5890</td>
</tr>
</tbody>
</table>

Table 2.8: Performances of confidence intervals: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.

- Subsampling: resample (with or without replacement) in $X$ a random sample of size $m = \lceil n^\gamma \rceil$ where $\gamma < 1$.

- Not easy to choose $\gamma$: Work still in progress...
2.6 Some Programs in MATLAB

2.6.1 Basic bootstrap functions

function xb=boot(x)

% written by L. SIMAR, Institute of Statistics,
% University of Louvain, Louvain-la-Neuve, Belgium
% function xb=boot(x)
% xb is a resample with replacement from a vector x
% x MUST BE A COLUMN VECTOR !!!!!!!!
% [n,k]=size(x);
if k>1
    disp('warning: x is not a column in function boot')
end
xb=x(floor(n*rand(size(x))+1));

function xb=resample(x,m)

% written by L. SIMAR, july 2001, Institute of Statistics,
% University of Louvain, Louvain-la-Neuve, Belgium
% function xb=resample(x,m)
% xb is a resample with replacement from a matrix x: (n x k)
% the entire ROW of x is drawn at each step
% m can be smaller, equal or larger than n
% [n,k]=size(x);
   sample=floor(n*rand(m,1)+1);
   xb=x(sample,:);
%
2.6.2 Example for $\bar{X}$ in $Exp(\mu)$

% Bootstrap on xbar for exponential law (Program written by L. SIMAR)
%  
% one sample of size n from expo(mu)
% where mu is the MEAN (usually 1 over lambda)
%  
% select B (bootstrap replications), mu and n
ngrid=100
B=5000
mu=50
n=50
  
x=exprnd(mu,n,1);
xbar=mean(x);s=std(x);
disp('smpl mean smpl std')
disp([xbar s])
xbarb=[];
for b=1:B
  xb=boot(x);
  mb=mean(xb);
  xbarb=[xbarb;mb];
end
xbarbar=mean(xbarb);sxbar=std(xbarb);
disp('xbarbar sxbar')
disp([xbarbar sxbar])

% kernel smoothing of the histogram: select the bandwith
s1=std(xbarb);r1=iqr(xbarb);h1=1.06*min([s1 r1/1.349])*B^(-1/5)
% bootstrap approx
[t1,f1]=Nkernel(xbarb,h1,ngrid);
plot(t1,f1,'-')
hold on
alf=n;bet=mu/n;
% true gamma
y=gampdf(t1,alf,bet);
plot(t1,y,'-.')
% normal approx with estimated parameters
y=normpdf(t1,xbar,s/sqrt(n));
plot(t1,y,'--')
hold off
2.6.3 Monte-Carlo performance evaluation

We present a code for the example of estimating $\exp(-1/\mu)$ in $X \sim \text{Expo}(\mu)$

```matlab
% Bootstrap on a function of xbar (Program written by L. SIMAR)
% MONTE-CARLO PERFORMANCES

% X is an Expo(mu)
%
% Parameter of interest: p=prob(NO event in one unit of time)=exp(-1/mu)
%
% MLE of p = exp(-1/xbar)
%
% generation of M samples of size n from expo(mu)
% where mu is the MEAN (usually 1 over lambda)
%
% Estimate the coverage probabilities
%
% Select B,M, mu and n
B=2000
mu=10
n=500
alpha=0.05
M=1000
p=exp(-1/mu); % this is the true value of \pi
%
t0 = clock;
MCnorm=[];MCboot=[];MCperc=[];MCtboot=[];
for mc=1:M
x=exprnd(mu,n,1); % this is a random sample of size n \Xs
xbar=mean(x);
s=std(x);
phat=exp(-1/xbar); % this is the MLE \hat \pi
pboot=[];studb=[];
for b=1:B
xb=boot(x); % this is a bootstrap sample \Xs^*(b)
mb=mean(xb);
pb=exp(-1/mb); % this \hat \pi^*(b)
ptb=sqrt(n)*mb*(pb-phat)/pb;
pboot=[pboot;pb];
studb=[studb;ptb];
end
[pboots,I] = sort(pboot);
```
[studbs,J]= sort(studb);
k1=floor(B*alpha./2)+1;
k2=floor(B*(1-(alpha./2)));
v1star=pboots(k1);
v2star=pboots(k2);
lowerc=v1star;
upperc=v2star;
lowboot=2*phat - v2star;
upboot=2*phat - v1star;
ylow=studbs(k1);
ysup=studbs(k2);

% Normal approximation: delta method p=exp(-1/mu)
% stdp=phat/(sqrt(n)*xbar);
z2=norminv(1-alpha/2,0,1);
lownorm=phat-z2*stdp;
upnorm =phat+z2*stdp;
%
% Bootstrap-t
%
lowtboot=phat-ysup*stdp;
uptboot=phat-ylow*stdp;

% MCnorm=[MCnorm;lownorm upnorm];
MCboot=[MCboot;lowboot upboot];
MCperc=[MCperc;lowerc upperc];
MCtboot=[MCtboot;lowtboot uptboot];

end

% t1=etime(clock,t0);
disp('elapsed time')
disp([t1])

% avlengnorm=mean(MCnorm(:,2)-MCnorm(:,1))
avlengboot=mean(MCboot(:,2)-MCboot(:,1))
avlenghtboot=mean(MCtboot(:,2)-MCtboot(:,1))
avlengperc=mean(MCperc(:,2)-MCperc(:,1))
covnorm=sum(MCnorm(:,1)<= p & p <= MCnorm(:,2))/M
covboot=sum(MCboot(:,1)<= p & p <= MCboot(:,2))/M
covtboot=sum(MCtboot(:,1)<= p & p <= MCtboot(:,2))/M
covperc=sum(MCperc(:,1)<= p & p <= MCperc(:,2))/M
save CIprop mu p n B M avleng* cov* MC*
### 2.6.4 Some useful smoothing functions for densities

```matlab
function [t,f]=Nkernel(y,h,ngrid)
% % written by L. SIMAR, Institute of Statistics, University of Louvain
% Louvain-la-Neuve, Belgium
% % function [t,f]=Nkernel(y,h,ngrid)
% Compute Gaussian Kernel density estimate from
% data vector y, bandwidth h for ngrid number of point
% between ymin - (range/6) and ymax + (range/6).
% Plot the obtained density if desired
% % OUTPUT is t=grid and f= value of density estimates
% %
% [n,p]=size(y);
% r=max(y)-min(y);
% t1=min(y)-r/6;t2=max(y)+r/6;
% pas=(t2-t1)/(ngrid-1);
% t=[t1:pas:t2]';
% a=(t*ones(1,n)-ones(ngrid,1)*y')/h;
% fa=normpdf2(a);
% f=sum(fa')/(n*h);

function f=normpdf2(x)
% % written by L. SIMAR, Institute of Statistics, University of Louvain
% Louvain-la-Neuve, Belgium
% % function f=normpdf2(x)
% compute the standard normal pdf for the elements
% of a matrix x
% f=exp((-0.5).*x.^2)./sqrt(2*pi);```
Chapter 3

The Properties of the Bootstrap

3.1 Why the Bootstrap works

• Consider a simple example where an asymptotical pivotal quantity is available
  – We have (MLE, or Delta Method, ...):
    \[ S = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\sigma(F)} \right) \sim AN(0, 1) \]
    where \( \sigma^2(F) = n\text{Var}_F(T(\mathcal{X})) \).
  – The confidence interval could be such that:
    \[ P \left[ T(\mathcal{X}) - z_{1-\alpha/2} \frac{\sigma(F)}{\sqrt{n}} \leq \theta(F) \leq T(\mathcal{X}) + z_{1-\alpha/2} \frac{\sigma(F)}{\sqrt{n}} \right] \approx 1 - \alpha \]

• (1) Suppose \( \sigma \) is known.
  – Edgeworth expansion (under regularity conditions):
    \[ Dist_{S,F}(x) = P_F(S \leq x) = \Phi(x) + n^{-\frac{1}{2}} p(x) \phi(x) + O(n^{-1}) \]
    where \( \Phi(\cdot) \) and \( \phi(\cdot) \) are the cdf and pdf of a standard normal, \( p(\cdot) \) is an even quadratic polynomial whose coefficients are cumulants of \( S \) such as skewness,...
– So we obtain:

$$\text{Dist}_{S,F}(x) - \Phi(x) = O(n^{-\frac{1}{2}})$$

**First Order Accuracy of the asymptotic approximation**

– In the bootstrap world, the bootstrap version of $S$ is:

$$S^* = \sqrt{n} \left( \frac{T(X^*) - \theta(F_n)}{\sigma(F_n)} \right)$$

where $\sigma(F_n) = \hat{\sigma}$ and $\theta(F_n)$ is usually $T(\mathcal{X})$.

– Edgeworth expansion for the bootstrap (under same regularity conditions):

$$\text{Dist}_{S}^*(x) = P_{F_n}(S^* \leq x) = \Phi(x) + n^{-\frac{1}{2}} \hat{p}(x)\phi(x) + O_p(n^{-1})$$

where $\hat{p}(\cdot)$ is the plug-in version of $p(\cdot)$. Now, typically,

$$p(x) - \hat{p}(x) = O_p(n^{-\frac{1}{2}})$$

– So we have here:

$$\text{Dist}_{S}^*(x) - \text{Dist}_{S,F}(x) = O_p(n^{-1})$$

**Second Order Accuracy of the bootstrap approximation**

– Similar argument will be valid for nonnormal limiting distribution, provided it does not depend on unknowns.

The bootstrap approximation works better than the asymptotic normal approximation (if $p(x) \neq 0$).
• (2) If $\sigma$ is unknown adn an estimator is available

- Suppose that an estimator $\hat{\sigma} = \hat{\sigma}(\mathcal{X})$ is available: we can studentize the root.
  \[
  U = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\hat{\sigma}(\mathcal{X})} \right) \sim AN(0, 1)
  \]

- The confidence interval here could be such that:
  \[
  P \left[ T(\mathcal{X}) - z_{1-\alpha/2} \frac{\hat{\sigma}(\mathcal{X})}{\sqrt{n}} \leq \theta(F) \leq T(\mathcal{X}) + z_{1-\alpha/2} \frac{\hat{\sigma}(\mathcal{X})}{\sqrt{n}} \right] \approx 1 - \alpha
  \]

- The Edgeworth expansion (under regularity conditions):
  \[
  Dist_{U,F}(x) = P_F(U \leq x) = \Phi(x) + n^{-\frac{1}{2}} q(x) \phi(x) + O(n^{-1})
  \]
  where $q(\cdot)$ is an even quadratic polynomial like $p(\cdot)$
  \[
  Dist_{U,F}(x) - \Phi(x) = O(n^{-\frac{1}{2}})
  \]

**First Order Accuracy of the asymptotic approximation**

- The Bootstrap version of $U$:
  \[
  U^* = \sqrt{n} \left( \frac{T(\mathcal{X}^*) - \theta(F_n)}{\hat{\sigma}(\mathcal{X}^*)} \right)
  \]
  where $\hat{\sigma}(\mathcal{X}^*) = \hat{\sigma}^*$ and , as often,$\theta(F_n) = T(\mathcal{X})$.

- The Edgeworth expansion (under regularity conditions):
  \[
  Dist_{U}(x) = P_{F_n}(U^* \leq x) = \Phi(x) + n^{-\frac{1}{2}} \hat{q}(x) \phi(x) + O_p(n^{-1})
  \]
  where $\hat{q}(\cdot)$ is the plug-in version of $q(\cdot)$. Usually again :
  \[
  q(x) - \hat{q}(x) = O_p(n^{-\frac{1}{2}})
  \]
– So we have:

\[ Dist^*_U(x) - Dist_{U,F}(x) = O_p(n^{-1}) \]

**Second Order Accuracy of the bootstrap approximation**

– \( Dist^*_U(x) \) can be used to construct bootstrap-t confidence intervals.

\[
\left[ T(\mathcal{X}) - n^{-1/2}\hat{\sigma}(\mathcal{X})u^*_{1-\alpha/2}, T(\mathcal{X}) - n^{-1/2}\hat{\sigma}(\mathcal{X})u^*_{\alpha/2} \right]
\]

where \( u^*(a) \) is the \( a \)-quantile of \( Dist^*_U(x) \).

– The Bootstrap-t is more accurate than the usual normal approximation (if \( q(x) \neq 0 \)).

• **(3) If \( \sigma(F) \) is unknown and no estimators are available**

  – we can not use the bootstrap-t, we only can use the basic bootstrap or the percentile bootstrap.

  – what about the accuracy? see next section.
3.2 The Virtue of prepivoting

• Consider the following root and its bootstrap version:

\[
W = \sqrt{n}(T(X) - \theta(F))
\]

\[
W^* = \sqrt{n}(T(X^*) - T(X))
\]

– As a matter of fact, we have:

\[
W = \sigma S = \sigma(F)S
\]

\[
W^* = \hat{\sigma} S^* = \sigma(F_n)S^*
\]

where \( S \) and \( S^* \) are defined in the preceding section.

– So, we have successively:

\[
\text{Dist}_{W,F}(x) = P_F(W \leq x) = P_F(S \leq \frac{x}{\sigma})
\]

\[= \Phi\left(\frac{x}{\sigma}\right) + n^{-\frac{1}{2}} p\left(\frac{x}{\sigma}\right) \phi\left(\frac{x}{\sigma}\right) + O(n^{-1})\]

\[
\text{Dist}_{W}^*(x) = P_{F_n}(W^* \leq x) = P_{F_n}(S^* \leq \frac{x}{\hat{\sigma}})
\]

\[= \Phi\left(\frac{x}{\hat{\sigma}}\right) + n^{-\frac{1}{2}} \hat{p}\left(\frac{x}{\hat{\sigma}}\right) \phi\left(\frac{x}{\hat{\sigma}}\right) + O_p(n^{-1})\]

– Now we have:

\[
\text{Dist}_{W}^*(x) - \text{Dist}_{W,F}(x) = \Phi\left(\frac{x}{\hat{\sigma}}\right) - \Phi\left(\frac{x}{\sigma}\right) + O_p(n^{-1})
\]

\[= (\hat{\sigma} - \sigma) R(x) + O_p(n^{-1})\]

– Since \((\hat{\sigma} - \sigma) = \sigma(F_n) - \sigma(F) = O_p(n^{-1/2})\), we obtain:

\[
\text{Dist}_{W}^*(x) - \text{Dist}_{W,F}(x) = O_p(n^{-\frac{1}{2}})
\]

First Order Accuracy of the bootstrap approximation
- It is always better to Studentize, if possible!
- It is not less accurate than the usual Normal approximation
  
  - NB1: Normal approximation is often unavailable;
  - NB2: Normal approximation is often useless if $\sigma$ is unknown and no estimators available.

- $\text{Dist}^*_W(x)$ can be used to construct confidence intervals for $\theta(F)$ (Basic bootstrap method):

  \[
  \left[ T(\mathcal{X}) - \frac{1}{\sqrt{n}} w^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) - \frac{1}{\sqrt{n}} w^*(\frac{\alpha}{2}) \right]
  \]

  where $w^*(a)$ is the $a$-quantile of $\text{Dist}^*_W(x)$.

- If appropriate, percentile or BC-percentile can also be used.

  \bullet\ Conclusion : the bootstrap may improve the approximation when (asymptotic) pivotal quantities are available.

  \bullet\ If such an estimator of $\sigma$ is not available : iterated bootstrap could be a solution (see below).
3.3 Consistency of the Bootstrap

- We consider here a general situation. We have a root $R_n(\mathcal{X}, F)$ and we are looking for an estimate of:

$$G_{F,n}(x) = \text{Dist}_{R_n,F}(x) = P_F(R_n \leq x)$$

- We suppose the root has been properly scaled such that $R_n(\mathcal{X}, F)$ has a nondegenerate limiting distribution $Q(x, F)$.

  - For example $R_n(\mathcal{X}, F) = n^{1/2}(\bar{X} - \mu) \sim AN(0, \sigma^2)$.
  
  - The bootstrap version of the root is $R^*_n = R_n(\mathcal{X}^*, F_n)$ for instance, $R^*_n = n^{1/2}(\bar{X}^* - \bar{X})$.

- The bootstrap estimate of $G_{F,n}(x)$ is

$$\hat{G}_{F,n}(x) = G_{F,n,n}(x) = \text{Dist}^*_{{R_n,F}}(x) = \text{Dist}_{R_n,F_n}(x) = P_{F_n}(R^*_n \leq x)$$

- Under mild regularity conditions, we often have an asymptotic expansion for $G_{F,n}(x)$:

$$\text{Dist}_{R_n,F}(x) = Q(x, F) + n^{-\frac{1}{2}}q_1(x, F) + n^{-1}q_2(x, F) + o(n^{-1})$$

where $Q$ is the asymptotic cdf of $R_n$, $q_1(x, F)$ is an even function of $x$ for each $F$ and $q_2(x, F)$ is an odd function of $x$ for each $F$.

- The bootstrap version for the distribution of $R^*_n$ is:

$$\text{Dist}^*_{{R_n,F}}(x) = Q(x, F_n) + n^{-\frac{1}{2}}q_1(x, F_n) + n^{-1}q_2(x, F_n) + o_p(n^{-1})$$

- Usually, for $i = 1, 2$, uniformly in $x$, we have:

$$q_i(x, F_n) - q_i(x, F) = O_p(n^{-\frac{1}{2}})$$
• So that we have:

$$\text{Dist}^*_{R_n}(x) - \text{Dist}_{R_n,F}(x) = Q(x, F_n) - Q(x, F) + O_p(n^{-1})$$

• The key of the bootstrap: the idea.

– if $Q$ is sufficiently “smooth” in $F$ (see below),

$$Q(x, F_n) - Q(x, F) = O_p(F_n(x) - F(x)) = O_p(n^{-\frac{1}{2}})$$

– so that the bootstrap works: we have the consistency!

$$\text{Dist}^*_{R_n}(x) - \text{Dist}_{R_n,F}(x) = O_p(n^{-\frac{1}{2}})$$

• With an appropriate chosen root it is even more accurate than the usual first order asymptotic approximation $Q(x, F)$.

– Suppose we have an asymptotic pivotal root:

$$Q(x, F) = Q(x) = Q(x, F_n)$$

- this is for example the case for

$$R_n(\mathcal{X}, F) = \sqrt{n} \frac{\bar{X} - \mu}{S} \sim AN(0, 1),$$

for all $F$ such that $\mu$ and $\sigma^2$ finite.

– In this case:

$$\text{Dist}^*_{R_n}(x) - \text{Dist}_{R_n,F}(x) = O_p(n^{-1})$$

Second order approximation for the bootstrap,

with (asymptotical) pivotal roots
• **Consistency of the Bootstrap: a general theorem**

Consider a neighbourhood $\mathcal{F}$ of $F$ in a suitable functional space of cdf. Suppose that $\mathcal{F}$ is such that:

$$P(F_n \in \mathcal{F}) \to 1 \quad \text{as } n \to \infty$$

If:

1. for all cdf $A \in \mathcal{F}$, $G_{A,n} = Dist_{R_n,A}$ converges weakly to $Q(\cdot, A)$:

$$\int g(x) \, dG_{A,n}(x) \to \int g(x) \, dQ(x, A), \quad \text{as } n \to \infty,$$

for all integrable function $g(\cdot)$;

2. the convergence is uniform on $\mathcal{F}$;

3. the mapping $A \to Q(\cdot, A)$ is continuous in $A$,

then

$$\forall x \text{ and } \forall \epsilon > 0, \quad P \left( |Dist_{R_n,F_n}(x) - Q(x, F)| > \epsilon \right) \to 0 \quad \text{as } n \to \infty.$$  

- The first condition ensures that there is a limit to converge to, for $G_{F,n}$ but also for $G_{F_n,n}$.

- As $n$ increases, $F_n$ changes. The second and third conditions are needed to ensure that $G_{F_n,n}(\cdot)$ approaches $Q(\cdot, F)$ along every possible sequence of $F_n$.

- All the 3 conditions are necessary to get a set of sufficient conditions for consistency.
• Example of a failure: $X_i \sim F(0, \theta)$, $i = 1, \ldots, n$.
  
  - Consider the root
    \[
    R_n(\mathcal{X}, F) = n \frac{\theta - X(n)}{\theta}
    \]
  where $X(n) = \max(X_1, \ldots, X_n)$
  
  - The bootstrap version is
    \[
    R_n^* = R_n(\mathcal{X}^*, F_n) = n \frac{X(n) - X^*(n)}{X(n)}
    \]
  where $X^*(n) = \max(X_1^*, \ldots, X_n^*)$
  
  - It is easily seen that
    \[
    P_{F_n}(R_n^* = 0) = 1 - (1 - \frac{1}{n})^n
    \]
    
    Therefore
    \[
    \lim_{n \to \infty} P_{F_n}(R_n^* = 0) = 1 - \exp(-1) = 0.6321
    \]
  
  - Whereas, generally, the asymptotic distribution of $R_n(\mathcal{X}, F)$ is continuous and so
    \[
    P_F(R_n = 0) = 0
    \]
  
  - Example, if $F$ is uniform $U(0, \theta)$, the asymptotic distribution of $R_n(\mathcal{X}, F)$ is
    \[
    Q(x, F) = Q(x) = \text{Expo}(1)
    \]
  
  - $\text{Dist}_{R_n, F_n}$ cannot converge to $Q(x, F)$, the standard exponential.
  
  - Here the second condition above fails: the distributional convergence for the root is not uniform on useful neighbourhoods of $F$. 
3.4 Some Useful Tools for the Asymptotics

As pointed above, we can see the problem as to analyze the asymptotic behavior of $Q(x, F_n) - Q(x, F)$. So we need some topological tools on space of cdf.

• The “Mallows metric”

- We will consider the following functional space:

  $$\Gamma_p = \left\{ G \mid \int |x|^p dG(x) < \infty \right\}.$$

- The “Mallows metric” defines a distance between $G$ and $H$ in $\Gamma_p$. It is defined by:

  $$d_p(G, H) = \inf_{C(X,Y)} \left( E(|X - Y|^p)^{\frac{1}{p}} \right)$$

  where $C(X,Y)$ is the set of pairs of random variables $(X, Y)$ having marginal cdf $G$ and $H$ respectively.

- Remark: $x$ could be multidimensional, then $|\cdot|$ denotes the usual euclidean norm $\| \cdot \|$.

- For random variable $X$ and $Y$ having distributions $G$ and $H$ respectively, $G$ and $H$ in $\Gamma_p$, we write with an abuse of notation:

  $$d_p(X, Y) = d_p(G, H),$$

  and we say, with an abuse of language:

  $$X \text{ and } Y \in \Gamma_p.$$
- **Basic properties**

  - (P1): Let $G_n, G \in \Gamma_p$, then $d_p(G_n, G) \to 0$ as $n \to \infty$ if and only if

    \[
    G_n \to G, \text{ weakly, and}
    \int |x|^p dG_n(x) \to \int |x|^p dG(x)
    \]

  - (P2): If $X_i, i = 1, \ldots, n$ is a random sample i.i.d. with cdf $F \in \Gamma_p$, we have:

    \[
    \text{as } n \to \infty, \ d_p(F_n, F) \to 0, \ a.s.
    \]

  - (P3): For any scalar $a$, and random variables $U$ and $V \in \Gamma_p$, we have:

    \[
    d_p(aU, aV) = |a| \ d_p(U, V).
    \]

  - (P4): For any random variables $U$ and $V \in \Gamma_2$, we have:

    \[
    [d_2(U, V)]^2 = [d_2(U - E(U), V - E(V))]^2 + |E(U) - E(V)|^2
    \]

  - (P5): For any two sequences of independent r.v. $U_i$ and $V_i$ in $\Gamma_p$, we have:

    \[
    d_p(\sum_{i=1}^{n} U_i, \sum_{i=1}^{n} V_i) \leq \sum_{i=1}^{n} d_p(U_i, V_i)
    \]

  - (P6): For any two sequences of independent r.v. $U_i$ and $V_i$ in $\Gamma_2$ with $E(U_i) = E(V_i)$, we have:

    \[
    \left[ d_2(\sum_{i=1}^{n} U_i, \sum_{i=1}^{n} V_i) \right]^2 \leq \sum_{i=1}^{n} [d_2(U_i, V_i)]^2
    \]
• **A simple example: bootstrapping a vector of mean.**
  
  Consider a random sample \( X_i \in \mathbb{R}^k \) with cdf \( F \in \Gamma_2 \). Let \( E(X) = \mu \). Define the root and its bootstrap analog:
  \[
  U_n = \sqrt{n}(\bar{X} - \mu) \quad \text{and} \quad U^*_n = \sqrt{n}(\bar{X}^* - \bar{X})
  \]
  
  Denote \( K(x) = P_F(U_n \leq x) \) and its bootstrap estimate \( \hat{K}(x) = P_{F_n}(U^*_n \leq x) \).
  
  We have successively:
  \[
  d_2(\hat{K}, K) = d_2(\sqrt{n}(\bar{X}^* - \bar{X})), \sqrt{n}(\bar{X} - \mu))
  \]
  \[
  = d_2 \left( \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X^*_i - \bar{X}), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \right)
  \]
  \[
  = \frac{1}{\sqrt{n}} d_2 \left( \sum_{i=1}^{n} (X^*_i - \bar{X}), \sum_{i=1}^{n} (X_i - \mu) \right) \quad \text{by [P3]}
  \]
  \[
  \leq \frac{1}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} \left[ d_2(X^*_i - \bar{X}, X_i - \mu) \right]^2} \quad \text{by [P6]}
  \]
  
  Since the sampling is iid the latter is equal to \( \frac{\sqrt{n(d_2(X^*_1 - X, X_1 - \mu))^2}}{\sqrt{n}} \):
  \[
  d_2(X^*_1 - \bar{X}, X_1 - \mu) = d_2(X^*_1 - E^*(X^*_1), X_1 - E(X_1))
  \]
  \[
  = \sqrt{[d_2(X^*_1, X_1)]^2 - \| E^*(X^*_1) - E(X_1) \|^2}
  \]
  \[
  = \sqrt{[d_2(F_n, F)]^2 - \| \bar{X} - \mu \|^2}
  \]
  
  Since the two terms in the last expression tends a.s. to zero when \( n \to \infty \), the \( d_2 \)-consistency of the bootstrap is proven:
  \[
  \text{as } n \to \infty, \ d_2(\hat{K}, K) \to 0, \text{ a.s.}
  \]
Chapter 4

Hypothesis Testing

4.1 Introduction

- We want to test an hypothesis on $F$: $H_0 : F = F_0$
  - Simple hypothesis: $F_0$ is fully specified (example: $F = N(0, 1)$)
  - Composite hypothesis: $F_0$ has some not determined aspects (example: $F_0 = N(0, \sigma^2)$).
  - We call the additional parameters $\eta$, the nuisance parameters (example: $\eta = \sigma^2$ if we want to test a fixed value of the mean of a normal variable)

- Test statistics:
  Suppose the test statistic is $T(\mathcal{X})$ and we reject $H_0$ if $T(\mathcal{X})$ is too large (or too small).
  - $p$-value of $H_0$
    - Let $t = T_{observed}$: we reject if $t$ is too large.
- The \( p \)-value or \textit{Attained Significance Level} is

\[
p = \Prob(T(X) \geq t \mid H_0).
\]

- Critical value for a test of size \( \alpha \):

- We want to size the type I error at the level \( \alpha \)
- Rejection Region: \( \{t \geq t_\alpha\} \) where \( t_\alpha \) such that:

\[
\Prob(T(X) \geq t_\alpha \mid H_0) = \alpha
\]

- How to compute \( p \)-value and/or \( t_\alpha \)?

• We need to approximate the sampling distribution of \( T(X) \) under \( H_0 \): Use of Monte-Carlo methods

  - \textbf{if} \( H_0 \) \textbf{is simple}: easy
    - either we know the sampling sampling distribution of \( T(X) \) under \( H_0 \)
    - either we simulate it by MC techniques:
      * simulation of iid random samples of size \( n \): \( X^* \) drawn from \( F_0 \)
      * take the Monte-Carlo empirical distribution of \( T(X^*) \) as approximation.
- if $H_0$ is composite: more complicated

- the $p$-value is not clearly defined because $\text{Prob}(T(\mathcal{X}) \geq t_\alpha | F \in H_0)$ may depend upon which $F$ satisfying $H_0$ is taken.

- in some case no problem: example the Student-$t$ test $T(\mathcal{X})$ has the same distribution for any $F$ verifying $H_0$

- The main idea of bootstrap test: estimate $p$-values by

$$\hat{p} = \text{Prob}(T(\mathcal{X}) \geq t | \hat{F}_0).$$

where $\hat{F}_0$ is a cdf which satisfies $H_0$.

- simulate $\mathcal{X}^{*b}$ under $\hat{F}_0$, for $b = 1, \ldots, B$

- Compute $t^{*}_b = T(\mathcal{X}^{*b}), b = 1, \ldots, B$

- We have $B+1$ values $t, t_1^{*}, \ldots, t_B^{*}$ under $H_0$, equally likely values for $T(\mathcal{X})$, so estimate of $p$:

$$\hat{p} = \frac{\#\{t^{*}_b \geq t\} + 1}{B + 1}$$

**Problem: How to choose $\hat{F}_0$?**

depends on the problem!
4.2 Parametric Bootstrap tests

• Suppose, in a parametric model, we are testing $H_0 : \theta = \theta_0$, in the presence of nuisance parameters $\eta$. We have:

$$F(x) = F(x|\theta, \eta)$$

• Suppose the rejection region is: $T(\mathcal{X})$ too large.

Computation of the $p$-value:

– How to define $\widehat{F}_0$?

  - Let $\hat{\eta}_0$ be the MLE of $\eta$ computed under the null $\theta = \theta_0$

    * Under regularity conditions, $\hat{\eta}_0$ converges in probability to a limit $\eta_0$ called the pseudo-true value of $\eta$

    - Then:

    $$\widehat{F}_0(x) = F(x|\theta_0, \hat{\eta}_0)$$

– We generate $\mathcal{X}^{*b}$ under $\widehat{F}_0$

– Finally:

$$\hat{p} = \frac{\#\{t_b^* \geq t\} + 1}{B + 1}$$

• Example: separate family test

  – We want to test

$$H_0 : f(x) = f_0(x|\zeta)$$
$$H_1 : f(x) = f_1(x|\zeta)$$
- Here $\eta = (\zeta, \xi)$ and $\theta = 0$ or $1$ as indicator of the model.

- The likelihood ratio statistics is

$$T(\mathcal{X}) = n^{-1} \log \left( \frac{\mathcal{L}_1(\hat{\xi})}{\mathcal{L}_0(\hat{\zeta})} \right)$$

where $\hat{\zeta}$ and $\hat{\xi}$ are the MLEs and $\mathcal{L}_i$ is the likelihood under $H_i$.

- We reject $H_0$ in favor of $H_1$ if $T(\mathcal{X})$ is too large.

- The parametric bootstrap algorithm:
  - We generate $B$ random samples $\mathcal{X}^*$ of size $n$ from $f_0(x|\hat{\zeta})$
  - For each sample we compute $\hat{\zeta}^*$ and $\hat{\xi}^*$ by maximizing the log likelihoods of the simulated samples $\mathcal{X}^*$:

$$\ell^*_0(\zeta) = \sum \log(f_0(x_i^*|\zeta))$$
$$\ell^*_1(\xi) = \sum \log(f_1(x_i^*|\xi))$$

  - We compute the simulated log likelihood ratio statistics:

$$t^* = n^{-1} \left\{ \ell^*_1(\hat{\xi}^*) - \ell^*_0(\hat{\zeta}^*) \right\}$$

  - The bootstrap distribution of $t^* = T(\mathcal{X}^*)$ approximates the distribution of $T(\mathcal{X})$ under the null $H_0$.

- The $p$-value is estimated by:

$$\hat{p} = \frac{\# \{ t_b^* \geq t \} + 1}{B + 1}$$

where $t = T_{\text{observed}}$. 
• **Numerical example:** Duration data (see above).

  - We want to test a **Gamma** pdf for $X$ against a **Lognormal** pdf:

    $$
    f_0(x|\zeta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} \\
    f_1(x|\xi) = \frac{1}{\sigma x} \phi \left( \frac{\log x - \mu}{\sigma} \right).
    $$

  - The MLE’s are $\hat{\zeta}$ and $\hat{\xi}$

    $$
    \hat{\zeta} = (\hat{\alpha}, \hat{\beta}) = (0.5798, 141.0716) \\
    \hat{\xi} = (\hat{\mu}, \hat{\sigma}^2) = (3.3333, 2.8540)
    $$

  - The likelihood ratio test statistics is (after some calculation):

    $$
    t = n^{-1} \left\{ \ell_1(\hat{\xi}) - \ell_0(\hat{\zeta}) \right\} \\
    = -\frac{1}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2} + \log(\Gamma(\hat{\alpha})) - \hat{\alpha}\hat{\mu} + \hat{\alpha} + \hat{\alpha} \log(\hat{\beta})
    $$

  - We observe $t = 0.0036$.

  - Bootstrap algorithm:

    - We generate $B = 1000$ iid sample $X^*$ of size $n = 10$ from a **Gamma**($\hat{\alpha}, \hat{\beta}$)
    - For each sample $X^*$ we compute $\hat{\zeta}^*$ and $\hat{\xi}^*$ and the value $t^*$
    - The estimate $p$-value is

      $$
      \hat{p} = \frac{\#\{t^* \geq t\} + 1}{B + 1} = 0.2717
      $$

  - We do not reject the Gamma hypothesis.
– If we want to test an **Exponential** against a **Gamma**, we have

\[
\begin{align*}
f_0(x|\mu) &= \frac{e^{-x/\mu}}{\mu} \\
f_1(x|\zeta) &= \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}
\end{align*}
\]

– The MLE’s are \(\hat{\zeta}\) and \(\hat{\mu}\)

\[
\hat{\zeta} = (\hat{\alpha}, \hat{\beta}) = (0.5798, 141.0716)
\]

\[
\hat{\mu} = 81.8000
\]

– The likelihood ratio test statistics is (after some calculation):

\[
t = n^{-1} \left\{ \ell_1(\hat{\zeta}) - \ell_0(\hat{\mu}) \right\} \\
= (\hat{\alpha} - 1) \left( \frac{1}{n} \sum \log(x_i) \right) + \log(\hat{\mu}) + 1 - \log(\Gamma(\hat{\alpha})) - \hat{\alpha} - \hat{\alpha} \log(\hat{\beta})
\]

– We observe \(t = 0.1241\).

– **Bootstrap algorithm:**

  - We generate \(B = 1000\) iid sample \(X^*\) of size \(n = 10\) from a \(Exp(\hat{\mu})\)
  
  - For each sample \(X^*\) we compute \(\hat{\zeta}^*\) and \(\hat{\mu}^*\) and the value \(t^*\)
  
  - The estimate \(p\)-value is \(\hat{p} = (#\{t^* \geq t\} + 1)/(B + 1) = 0.1440\)

  – We do not reject the Exponential hypothesis (\(n = 10\): no clear definite answer).

– Here, we have a nested test (\(H_0\) is a restriction of \(H_1\)): the Chi-square test (asymptotic!) can be used:

\[
\text{If } H_0 \text{ is true } 2 \left\{ \ell_1(\hat{\zeta}) - \ell_0(\hat{\mu}) \right\} \sim \chi^2_{(1)} \text{ when } n \to \infty.
\]

\[
p\text{-value} = P(\chi^2_{(1)} \geq 0.1241*2*10) = P(\chi^2_{(1)} \geq 2.4818) = 0.1152.
\]
4.3 Nonparametric Permutation Tests

- **An old idea** (R.A. Fisher in the 1930’s)
  
  - No particular form assumed for data distribution
  
  - But how to define $\hat{F}_0$ in this nonparametric setup?
    
    - One solution: try to choose $T(\mathcal{X})$ such that $T(\mathcal{X})$ is distribution-free under $H_0$ (example: nonparametric rank tests, see *e.g.* Gibbons 1971)
    
    - Other solution, more widely applicable: eliminate the remaining unknown parameters when $H_0$ is true by conditioning on a sufficient statistics under $H_0$.
    
    - If $S$ is a sufficient statistics under the $H_0$, the conditional $p$-value may be computed as:
      
      $$p = P \left( T(\mathcal{X}) \geq t \mid H_0, S = s \right).$$

- **Permutation tests**: comparative tests (involving two sets of random variables) when under the null $H_0$, empirical cdfs are sufficient statistics.
  
  - In standard one sample models, the cdf $F_n$ or, equivalently, the order statistics $(X_{(1)}, \ldots, X_{(n)})$ is a sufficient statistics
  
  - In permutation comparative tests, $S$ will be appropriate empirical cdfs or, equivalently, appropriate order statistics, depending on the problem.
- Permutations tests are particularly useful as confirmatory test procedure when we have a “candidate” test statistics corresponding to some plausible parametric model: some $T(\mathcal{X})$ with a reasonable rejection region. Example, for comparison of two treatments: difference of means, or of medians, etc...

- **Example 1: The two-sample problem**

  - We have two independent samples $\mathcal{Y} = \{y_1, \ldots, y_m\}$ and $\mathcal{Z} = \{z_1, \ldots, z_n\}$ drawn from possibly different DGP, $F_Y$ ans $F_Z$. Let $F_Y$ be the control distribution and $Z$ be the treatment distribution.

  - We want to test $H_0 : F_Y = F_Z$

    - Suppose we observe $\hat{\theta}_{observed} = \bar{z} - \bar{y} \gg 0$, we expect that $H_0$ could be wrong and that the treatment has a positive effect.

    - We reject “$H_0 : no \ space difference$” in favor of “$H_1 : positive \ space effect$” if $\hat{\theta}$ is too large.

    - the $p$-value is

      $$p - value = P(\hat{\theta} > \hat{\theta}_{observed}|H_0).$$

    - The **parametric** approach, $F_Y = N(\mu_Y, \sigma^2)$ and $F_Z = N(\mu_Z, \sigma^2)$ leads to the Student-t test.

      $$p - value = P(t_{n+m-2} > t_{obs}),$$
with

$$t = \frac{\bar{z} - \bar{y}}{s_p \sqrt{1/n + 1/m}}.$$  

where $s_p^2 = ((n - 1)s^2_z + (m - 1)s^2_y)/(n + m - 2)$.

- In nonparametric models, for evaluating $P(\hat{\theta} \geq \hat{\theta}_{\text{observed}} | H_0)$ we don’t have the trick of the Student’s method to solve the problem: permutation test is one solution (bootstrap will be another below).

- **Permutation test for the two-sample problem:**

  - If the null is true, each observation $z_i$ and/or $y_i$ could have come from either distribution.

  - We combine the $N = m + n$ observation in a single sample $X = (Y, Z)$

  - Under the null, the sufficient statistics is the empirical cdf $F_N(x)$, or equivalently, the order statistics for the pooled sample $s = (x_{(1)}, \ldots, x_{(N)})$.

  - The $p$-value of $H_0$ is thus given by

    $$p - \text{value} = P(\hat{\theta}(X) > \hat{\theta}_{\text{observed}} | S = s, H_0)$$

  - Now, given that $S = s$, a random sample $X$ can only be a random permutation of $s$. The first $m$ elements forming $Y$ will be obtained by taking a random sample of size $m$ without replacement from $X$, the second sample $Z$ will be formed by the $n$ remaining ones. This is a random permutation of the sample: denoted by $X^*$.  

- **Permutation property:** Under $H_0$, all the $\binom{N}{m}$ permutation samples $\mathcal{X}^*$ are equally likely.

- The $p$-value of $H_0$ is now given by

$$p - \text{value} = P(\hat{\theta}(\mathcal{X}) > \hat{\theta}_{\text{observed}} \mid S = s, H_0)$$

$$= P(\hat{\theta}(\mathcal{X}^*) > \hat{\theta}_{\text{observed}})$$

$$= \frac{\text{# of permutations such that } \hat{\theta}(\mathcal{X}^*) \geq \hat{\theta}_{\text{observed}}}{\binom{N}{m}}$$

where $\hat{\theta}(\mathcal{X}^*)$ is a permutation replication of $\hat{\theta}(\mathcal{X})$.

- In practice we simulate $B$ random permutations of $(x_1, \ldots, x_N)$ and calculate $\hat{\theta}(\mathcal{X}^{*(b)})$, $b = 1, \ldots, B$ and approximate the $p$-value by:

$$\hat{p} = \frac{1 + \#\{\hat{\theta}(\mathcal{X}^{*(b)}) \geq \hat{\theta}_{\text{observed}}\}}{B + 1}.$$ 

- **Numerical illustration**

  - **Case 1:** Two independent samples drawn from normal populations

    $F_Y = N(10, (1.5)^2)$, sample size $n = 7$

    $F_Z = N(12, (1.5)^2)$, sample size $m = 10$

    - we observe the following statistics

      $$\bar{y} = 10.0236 \quad s_y^2 = 1.7213$$

      $$\bar{z} = 11.3601 \quad s_z^2 = 1.1464$$
- The student statistics \( s_p^2 = 1.3763 \) and \( \hat{\theta}_{\text{obs}} = 1.3365 \) takes the value \( t_{\text{obs}} = 2.3116 \).

- The (exact here) \( p \)-value is \( P(t_{15} \geq t_{\text{obs}}) = 0.0177 \)

- The permutation test on \( \hat{\theta} = \bar{Z} - \bar{Y} \) over \( B = 1000 \) simulation provides

\[
\hat{p} = \frac{1 + \# \{ \hat{\theta}(\mathbf{X}^{*}(b)) \geq 1.3365 \}}{B + 1} = 0.0105
\]

- Figure 4.1 illustrates the simulated empirical density of the \( B = 1000 \) values of \( \hat{\theta}(\mathbf{X}^{*}(b)) \).

Figure 4.1: Permutation simulated estimates of the sampling distribution of \( \hat{\theta}(\mathbf{X}^{*}) = \bar{z}^{*} - \bar{y}^{*} \).
- **Case 2**: Two independent samples of survival times from unknown populations

  - we have the samples $m = 10$ and $n = 7$:

    \[
    z = (2.5884 \ 1.4106 \ 8.6561 \ 1.4820 \ 26.1792 \ 0.7062 \ 0.7625 \ 1.025 \ 4.0447 \ 15.6797)
    \]

    \[
    y = (0.4306 \ 0.1853 \ 0.2734 \ 0.3542 \ 4.7347 \ 0.1250 \ 0.4149)
    \]

  - we have the following statistics

    \[
    \bar{y} = 0.9312 \quad s_y^2 = 2.8260
    \]

    \[
    \bar{z} = 6.3535 \quad s_z^2 = 71.2146
    \]

  - The student statistics ($s_p^2 = 43.8591$ and $\hat{\theta}_{obs} = 5.4223$) takes the value $t_{obs} = 1.6614$.

  - The $p$-value is $P(t_{15} \geq t_{obs}) = 0.0587$

  - The permutation test on $\hat{\theta} = \bar{Z} - \bar{Y}$ over $B = 1000$ simulation provides

    \[
    \hat{p} = \frac{1 + \#\{\hat{\theta}(X^{*\{b\}}) \geq 5.4223\}}{B + 1} = 0.0300
    \]

  - Figure 4.2 illustrates the simulated empirical density of the $B = 1000$ values of $\hat{\theta}(X^{*\{b\}})$.

  - Here the normal theory with equal variance is not very appropriate.
Figure 4.2: Permutation simulated estimates of the sampling distribution of $\hat{\theta}(X^*) = \bar{z}^* - \bar{y}^*$.

- **Remark**

Any other test statistics could be used with the permutation test: difference of trimmed means (robustness to outliers), difference of medians, etc...
• **Example 2: The Permutation test for Correlation**

Suppose we have \( X_i = (Y_i, Z_i) \) is a random sample of \( n \) pairs. We want to test the independence between \( Y \) and \( Z \).

- if the observed value of the empirical correlation is large (\( \hat{\rho} \gg 0 \)), we will be tempted to reject the null. The \( p \)-value is then \( P(\hat{\rho} \geq \hat{\rho}_{obs} | H_0) \) (one-sided case).

- **In general**, the bivariate \( F_n(x) : F_n(y, z) \) is the minimal sufficient statistics for \( F \)

- **Under the null**, \( F = F_Y F_Z \), the minimal sufficient statistics \( S \) is the two sets of orders statistics \((Y(1), \ldots, Y(n))\) and \((Z(1), \ldots, Z(n))\) (or, equivalently, the two empirical marginal cdfs \( F_{Y,n} \) and \( F_{Z,n} \)).

- The observed value of \( s \) is:

\[
  s = (y(1), \ldots, y(n), z(1), \ldots, z(n))
\]

- Under the null and under the constraint that \( S = s \), the random sample \( \mathcal{X} = \{X_i = (Y_i, Z_i) | i = 1, \ldots, n \} \) is equivalent to the random sample

\[
  \mathcal{X}^* = \{(y(1), Z_1^*), \ldots, (y(n), Z_n^*)\},
\]

where \( Z_1^*, \ldots, Z_n^* \) is a random permutation of \((z(1), \ldots, z(n))\).
- There are $n!$ such equally likely random permutations, so

$$p\text{-value} = P(\hat{\rho}(X) \geq \hat{\rho}_{\text{observed}} \mid S = s, H_0)$$

$$= P(\hat{\rho}(X^*) \geq \hat{\rho}_{\text{observed}})$$

$$= \frac{\# \text{ of permutations such that } \hat{\rho}(X^*) \geq \hat{\rho}_{\text{observed}}}{n!}$$

- In practice we simulate $B$ random permutations of $(z(1), \ldots, z(n))$ and calculate $\hat{\rho}(X^{*(b)})$, $b = 1, \ldots, B$ and approximate the $p$-value by:

$$\hat{p} = \frac{1 + \#\{\hat{\rho}(X^{*(b)}) \geq \hat{\rho}_{\text{observed}}\}}{B + 1}.$$

**Numerical illustration**

- **Case 1**: Two independent samples ($n = 20$) drawn from two normal populations ($\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = 2$ and $\sigma_2 = 1$). The obtained sample is plotted in Figure 4.3. We obtain the following results from the permutation test ($B = 1000$):

$$\hat{\rho} = 0.0929$$

$$\hat{p} = 0.3337$$

The right panel of Figure 4.3 provides the sampling distribution of $\hat{\rho}(X^*)$ under $H_0$. 
Figure 4.3: Two independent normal variates. Left panel: a simulated sample \((n = 20)\). Right panel: simulated estimates of the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\). Here, \(\hat{\rho} = 0.0929\)

- **Case 2**: A random sample \((n = 20)\) from a dependent bivariate normal with \(\mu_1 = 0, \mu_2 = 2, \sigma_1 = 2, \sigma_2 = 1\) and \(\rho = 0.50\). The obtained sample is plotted in Figure 4.4. We obtain the following results from the permutation test \((B = 1000)\):

\[
\hat{\rho} = 0.7205 \\
\hat{p} = 0.0010
\]

The right panel of Figure 4.4 provides the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\).
- **Case 3**: A random sample \((n = 100)\) from dependent bivariate normal with \(\sigma_1 = 2, \sigma_2 = 1\) and \(\rho = 0.25\). The obtained sample is plotted in Figure 4.5. We obtain the following results from the permutation test \((B = 1000)\):

\[
\hat{\rho} = 0.2560 \\
\hat{p} = 0.0050
\]

The right panel of Figure 4.5 provides the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\).
Figure 4.5: A random sample from bivariate normal with $\rho = 0.25$. Left panel: a simulated sample ($n = 100$). Right panel: simulated estimates of the sampling distribution of $\hat{\rho}(X^*)$ under $H_0$. Here, $\hat{\rho} = 0.2506$. 
4.4 Nonparametric Bootstrap Tests

• The same idea as nonparametric permutation tests but resampling with replacement and not without replacement, but can be applied to much wider class of problems.

  - We must resample from a \( \hat{F}_0 \) which satisfies \( H_0 \): it cannot be simply \( F_n \).

    - Depends on the problem
    - Many candidates here for \( \hat{F}_0 \), depending on the restrictions imposed in addition to \( H_0 \).

  - Then the procedure is the same. If the rejection region is of the from \( T(\mathcal{X}) \) too large, the \( p \)-value obtained by the bootstrap is:

    \[
    \hat{p} = P( T(\mathcal{X}^*) \geq T_{\text{observed}} \mid \hat{F}_0 ) \\
    \approx 1 + \# \{ T(\mathcal{X}^{*b}) \geq T_{\text{observed}} \} / B + 1
    \]

• Example 1: Two-sample test

  - As above, we can choose many different statistics

    - Difference of the means \( \hat{\theta} = \bar{z} - \bar{y} \)

    - We can studentize it (same variance):

      \[
      t = \frac{\bar{z} - \bar{y}}{s_p \sqrt{1/n + 1/m}},
      \]
- We could also use (different variance):

\[ t = \frac{\bar{z} - \bar{y}}{\sqrt{s^2_Z/n + s^2_Y/m}}, \]

- \( F_N \), the empirical cdf of \( X \), is a natural estimator of the common cdf of \( Y \) and \( Z \).

- Numerical illustration:

  - 2 independent samples from unknown populations (see above)

  \[
  \begin{align*}
  z &= \{2.5884 \ 1.4106 \ 8.6561 \ 1.4820 \ 26.1792 \ 0.7062 \ 0.7625 \ 0.1250 \ 0.4149 \} \\
  y &= \{0.4306 \ 0.1853 \ 0.2734 \ 0.3542 \ 4.7347 \ 0.1250 \ 0.4149 \}
  \end{align*}
  \]

  - we had the following statistics

  \[
  \begin{align*}
  \bar{y} &= 0.9312 \quad & s^2_y &= 2.8260 \\
  \bar{z} &= 6.3535 \quad & s^2_z &= 71.2146 \\
  \hat{\theta}_{\text{obs}} &= 5.4223
  \end{align*}
  \]

  - The permutation test on \( \hat{\theta} = \bar{Z} - \bar{Y} \) over \( B = 1000 \) simulations provided \( \hat{p} = 0.0300 \).

  - The bootstrap test, with \( B = 1000 \) replications, provides

  \[
  \hat{p} = \frac{1 + \# \{ \hat{\theta}(X^{* (b)}) \geq 5.4223 \}}{B + 1} = 0.0390
  \]

  - Figure 4.6 illustrates the simulated empirical density of the \( B = 1000 \) values of \( \hat{\theta}(X^{* (b)}) \).
Extension: testing the equality of the density of two populations $H_0 : f_Z = f_Y$: nonparametric bootstrap test based on kernel density estimates. Test statistics is an estimation of the integrated squared error between the two densities

$$ISE = \int (f_Z(u) - f_Y(u))^2 \, du$$

see Li (1996, 1999).

**Example 2: One-sample test**

- We want to test if $\mu = \mu_0$ in an unknown population from a sample of size $n$. The alternative is $\mu > \mu_0$.
- We can use many statistics: we choose here

$$T(x) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

- How to select $\hat{F}_0$?
- It cannot be $F_n$: it does not obey $H_0$.

- We **shift** the empirical cdf so that it has mean $\mu_0$

\[ \tilde{x}_i = x_i - \bar{x} + \mu_0 \]

$\hat{F}_0$ is the empirical cdf of the $\tilde{x}$’s.

- The bootstrap test statistics is then

\[ T(\tilde{\mathbf{X}}^*) = \frac{\bar{\tilde{X}}^* - \mu_0}{\hat{S}^*/\sqrt{n}} \]

where $\bar{\tilde{X}}^*$ and $\hat{S}^*$ are the sample mean and std deviation of the bootstrap samples $\tilde{\mathbf{X}}^*$ obtained by drawing with replacement in the shifted values $\tilde{x}$.

- The p-value is then obtained by

\[ \hat{p} = P(T(\tilde{\mathbf{X}}^*) \geq T_{observed} \mid \hat{F}_0) \]

- Note that in this case, the bootstrap statistics can also be calculated without explicitly shifting:

\[ T(\tilde{\mathbf{X}}^*) \equiv \frac{\bar{\tilde{X}}^* - \bar{X}}{S^*/\sqrt{n}} \]

where $\bar{\tilde{X}}^*$ and $S^*$ are the sample mean and std deviation of the usual bootstrap samples $\tilde{\mathbf{X}}^*$. 
• Example 3: Test of independence

- Suppose we have $X_i = (Y_i, Z_i)$ is a random sample of $n$ pairs. We want to test the independence between $Y$ and $Z$.

  - if the observed value of the empirical correlation is large ($\hat{\rho} \gg 0$), we will be tempted to reject the null. The $p$-value is then $P(\hat{\rho}(X) \geq \hat{\rho}_{obs}|H_0)$ (one-sided case).

  - The bootstrap estimate of the $p$-value is obtained as

    $$\hat{p} = P(\hat{\rho}(X^*) \geq \hat{\rho}_{obs}|H_0)$$

    where $X^* = \{Y^*, Z^*\}$ with $Y^*$ and $Z^*$ being independent bootstrap samples from $Y$ and $Z$ respectively.

  - In practice we simulate $B$ bivariate random samples $X^{*\{b\}}$ and calculate $\hat{\rho}(X^{*\{b\}}), b = 1, \ldots, B$ and approximate the $p$-value by:

    $$\hat{p} = \frac{1 + \#\{\hat{\rho}(X^{*\{b\}}) \geq \hat{\rho}_{observed}\}}{B + 1}.$$ 

    - Poor power of the test is expected, unless bivariate normal population.
• Numerical illustration (as above)

- **Case 1**: Two independent samples \((n = 20)\) drawn from two normal populations \((\mu_1 = 0, \mu_2 = 2, \sigma_1 = 2 \text{ and } \sigma_2 = 1)\). The obtained sample is plotted in Figure 4.7. We obtain the following results from the bootstrap test \((B = 5000)\):

\[
\hat{\rho} = 0.1584 \\
\hat{p} = 0.2490
\]

The right panel of Figure 4.7 provides the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\).

![Figure 4.7: Two independent normal variates. Left panel: a simulated sample \((n = 20)\). Right panel: simulated estimates of the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\). Here, \(\hat{\rho} = 0.1584\) ]
– **Case 2**: A random sample \((n = 20)\) from a bivariate normal with \(\mu_1 = 0, \mu_2 = 2, \sigma_1 = 2, \sigma_2 = 1\) and \(\rho = 0.50\). The obtained sample is plotted in Figure 4.8. We obtain the following results from the bootstrap test \((B = 5000)\):

\[
\hat{\rho} = 0.7823 \\
\hat{p} = 0.0002
\]

The right panel of Figure 4.8 provides the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\).

![Figure 4.8: A random sample from bivariate normal with \(\rho = 0.50\). Left panel: a simulated sample \((n = 20)\). Right panel: simulated estimates of the sampling distribution of \(\hat{\rho}(X^*)\) under \(H_0\). Here, \(\hat{\rho} = 0.7823\)](image)

– **Case 3**: A random sample \((n = 100)\) from bivariate normal with \(\sigma_1 = 2, \sigma_2 = 1\) and \(\rho = 0.25\). The obtained sample is plotted in Figure 4.5. We obtain the following results from the bootstrap test \((B = 5000)\):

\[
\hat{\rho} = 0.1820 \\
\hat{p} = 0.0366
\]
The right panel of Figure 4.10 provides the sampling distribution of \( \hat{\rho}(\mathcal{X}^*) \) under \( H_0 \).

![Figure 4.9: A random sample from bivariate normal with \( \rho = 0.25 \). Left panel: a simulated sample \((n = 100)\). Right panel: simulated estimates of the sampling distribution of \( \hat{\rho}(\mathcal{X}^*) \) under \( H_0 \). Here, \( \hat{\rho} = 0.1820 \).](image1)

- **Case 4:** \( Y \sim N(0, 1) \) and \( Z = Y^2 \), so that \( \rho = 0 \).
  - We obtained with \( n=100 \), \( \hat{\rho} = 0.0953 \) and \( \hat{p} = 0.1622 \): non reject \( H_0 : \rho = 0 \).
  - Not powerful test of independence here.

![Figure 4.10: Left panel: a simulated sample \((n = 100)\). Right panel: simulated estimates of the sampling distribution of \( \hat{\rho}(\mathcal{X}^*) \) under \( H_0 \). Here, \( \hat{\rho} = 0.0953 \).](image2)
Chapter 5

The Bootstrap in Regression Models

5.1 Regression Models

• Consider the following linear regression model:

\[ y = X \beta + \epsilon \]

where \( y : (n \times 1) \), \( X : (n \times p) \), \( \beta : (p \times 1) \) and \( \epsilon : (n \times 1) \).

Element by element we have:

\[ y_i = x_i \beta + \epsilon_i \]

where \( x_i : (1 \times p) \) is the \( i \)th row of \( X \) (the first element being a one is the model has an intercept).

• The main assumptions are as follows:

  – The matrix \( X \) is not random
  
  – The \( \epsilon_i \) are i.i.d. with common cdf \( F \) with mean 0 and constant variance \( \sigma^2 \):

\[ \epsilon_i \sim F(0, \sigma^2) \]
– In matrix notation:
\[ \epsilon \sim (0, \sigma^2 I_n) \]

– As \( n \to \infty \),
\[ \frac{1}{n} X'X \to V > 0 \]

• Statistical analysis.

– OLS estimators:
\[
\hat{\beta} = (X'X)^{-1}X'y \\
\hat{\sigma}^2 = \frac{1}{n - p} (y - X\hat{\beta})' (y - X\hat{\beta})
\]

– The asymptotic:
\[
\sqrt{n}(\hat{\beta} - \beta) \to N_p(0, \sigma^2V^{-1}).
\]

– In particular we obtain the multivariate asymptotic root:
\[
(X'X)^{1/2} \frac{\hat{\beta} - \beta}{\sigma} \sim AN_p(0, I_p)
\]

– Element by element, we have a Studentized asymptotic pivotal root for \( \beta_j \), for \( j = 1, \ldots, p \):
\[
U_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{c_{jj}}} \sim AN(0, 1)
\]

where \( c_{jj} \) is the \( (j, j) \)th element of the matrix \( (X'X)^{-1} \).

– Asymptotic confidence intervals:
\[
\left[ \hat{\beta}_j - \hat{\sigma} \sqrt{c_{jj}} z_{1-\alpha/2}, \hat{\beta}_j - \hat{\sigma} \sqrt{c_{jj}} z_{\alpha/2} \right]
\]
• Properties of Residuals:

\[ \hat{\epsilon} = y - X\hat{\beta} \]
\[ = y - X(X'X)^{-1}X'y \]
\[ = M_X y \]

where \( M_X = I_n - X(X'X)^{-1}X' = I_n - H_X \).

- \( H_X \) is the hat matrix. Note that:

\[ H_X X = X \]
\[ M_X X = 0 \]

- Moments of \( \epsilon \):

\[ E_F(\hat{\epsilon}) = 0 \]
\[ Var_F(\hat{\epsilon}) = M_X(\sigma^2 I_n)M_X = \sigma^2 M_X = \sigma^2(I_n - H_X) \]

- So that, element by element, we have:

\[ E_F(\hat{\epsilon}_i) = 0 \]
\[ Var_F(\hat{\epsilon}_i) = (1 - h_i)\sigma^2 \]

where \( h_i \) is the \((i, i)\)th element of \( H_X \): \( h_i = x_i(X'X)^{-1}x_i' \).
5.2 Bootstrapping the Residuals

- In the **real world** the data generating process $\mathcal{P}$ is characterized by

$$\mathcal{P} = (\beta, F)$$

then, for a given $x_i$,

$$y_i = x_i \beta + \epsilon_i, \text{ where } \epsilon_i \sim F(0, \sigma^2).$$

This provides the original sample: $\mathcal{X} = \{(x_i, y_i) | i = 1, \ldots, n\}$.

- In the **bootstrap world** the data generating process is

$$\hat{\mathcal{P}} = (\hat{\beta}, F_n)$$

where $F_n$ is either the empirical cdf of the centered residuals (non-parametric bootstrap) or $F(0, \hat{\sigma}^2)$, if $F$ is a given parametric cdf (parametric bootstrap).

- The (nonparametric) bootstrap algorithm is implemented as follows:

  - If the model has no intercept, $E_{F_n}(\hat{\epsilon}_i) \neq 0$ without centering the residuals, $\Rightarrow$ center the residuals:

$$\hat{\epsilon}_i^{(1)} = \hat{\epsilon}_i - \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}_i$$

  - We define here $F_n$ as the empirical cdf of $\hat{\epsilon}_i^{(1)}$, $i = 1, \ldots, n$.

  - Draw independently from $F_n$ to produce the bootstrap residuals $\epsilon_i^*, i = 1, \ldots, n$. 
– Generate with $\hat{P}$ the bootstrap values:

$$y_i^* = x_i\hat{\beta} + \epsilon_i^*, \quad i = 1, \ldots, n$$

– The pseudo-sample $X^* = \{(x_i, y_i^*) | i = 1, \ldots, n\}$ provides the bootstrap value of $\hat{\beta}$:

$$\hat{\beta}^* = (X'X)^{-1}X'y^*$$

– The bootstrap analog of the root for $\beta_j$ is:

$$U_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{\hat{\sigma}^* \sqrt{c_{jj}}}$$

where $c_{jj}$ is the $(j, j)$th element of the matrix $(X'X)^{-1}$ and

$$\hat{\sigma}^* = \frac{1}{n - p} \frac{1}{(y^* - X\hat{\beta}^*)'(y^* - X\hat{\beta}^*)}.\$$

– In particular, the bootstrap-$t$ confidence interval for $\beta_j$ is:

$$\left[ \hat{\beta}_j - \hat{\sigma} \sqrt{c_{jj}} u_{1-\alpha/2}^*, \hat{\beta}_j - \hat{\sigma} \sqrt{c_{jj}} u_{\alpha/2}^* \right]$$

where $u_a^*$ is the $a-$quantile of $\text{Dist}_{U_j}^*(x)$.

• It can be proven that the bootstrap-$t$ achieve 3rd order accuracy for the slopes parameters ($O_p(n^{-3/2})$).
• Additional refinement:
  – Since $\text{Var}_F(\hat{\epsilon}_i) = (1 - h_i)\sigma^2 \neq \sigma^2$, we may also correct for the second moment and then center the obtained residuals.
  – Define the corrected residuals:
    \[
    \hat{\epsilon}_i^{(2)} = \frac{\hat{\epsilon}_i}{\sqrt{1-h_i}} - \sum_{j=1}^{n} \frac{\hat{\epsilon}_j}{\sqrt{1-h_j}}
    \]
  – Now we define $F_n$ as the empirical cdf of $\hat{\epsilon}_i^{(2)}$, $i = 1, \ldots, n$.

5.3 Monte-Carlo Performances of Confidence Intervals

• We simulated data according the following model:
  – The regression model is
    \[y_i = 2 + x_i + \epsilon_i\]
  – The error term is a shifted gamma distribution with mean zero:
    \[\epsilon_i \sim \text{Gamma}(a, c) - \mu\]
    where $\mu = ac$ and $\sigma^2 = ac^2$.
  – We choose $a = 2$ and $c = 1/\sqrt{2}$ to get $\sigma = 1$.
  – We choose a fixed design for $x_i$:
    \[x_i = (1/n, 2/n, \ldots, n/n)\]

• We compare the performances of the asymptotic normal method, the percentile method and the bootstrap-$t$ method.
We choose $M = 1000$ Monte-Carlo replications: so, in Table 5.1, the coverage probabilities of nominal level 0.95 with an error bound of 0.0138 (see Table 2.3).

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 5$</th>
<th></th>
<th>$n = 10$</th>
<th></th>
<th>$n = 20$</th>
<th></th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
</tr>
<tr>
<td>As. Normal</td>
<td>5.505</td>
<td>0.855</td>
<td>4.127</td>
<td>0.918</td>
<td>2.917</td>
<td>0.931</td>
<td>1.350</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>4.210</td>
<td>0.763</td>
<td>3.695</td>
<td>0.881</td>
<td>2.763</td>
<td>0.913</td>
<td>1.337</td>
</tr>
<tr>
<td>Bootstrap-t</td>
<td>8.620</td>
<td>0.941</td>
<td>4.824</td>
<td>0.953</td>
<td>3.096</td>
<td>0.943</td>
<td>1.363</td>
</tr>
</tbody>
</table>

Table 5.1: Performances of confidence intervals for $\beta_2 = 1$, here $\sigma = 1$: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.

- Same scenario as above with refinements for the bootstrap:
  - we bootstrap on the residuals $\tilde{\epsilon}_i^{(2)}$, $i = 1, \ldots, n$. The results are in Table 5.2.

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 5$</th>
<th></th>
<th>$n = 10$</th>
<th></th>
<th>$n = 20$</th>
<th></th>
<th>$n = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
</tr>
<tr>
<td>As. Normal</td>
<td>5.350</td>
<td>0.854</td>
<td>4.104</td>
<td>0.909</td>
<td>2.964</td>
<td>0.940</td>
<td>1.346</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>5.232</td>
<td>0.849</td>
<td>4.106</td>
<td>0.905</td>
<td>2.967</td>
<td>0.941</td>
<td>1.344</td>
</tr>
<tr>
<td>Bootstrap-t</td>
<td>8.424</td>
<td>0.954</td>
<td>4.794</td>
<td>0.946</td>
<td>3.158</td>
<td>0.960</td>
<td>1.358</td>
</tr>
</tbody>
</table>

Table 5.2: Performances of confidence intervals for $\beta_2 = 1$, Refined bootstrap algorithm for correcting the variance of the residuals. Here $\sigma = 1$: average length of the intervals and estimation of coverage probabilities (nominal level 95%) over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.

- The improvements are only significant for small sample sizes and especially for the percentile method.
– Table 5.3 gives an idea of the sampling variation of the correction terms $\sqrt{1 - h_i}$: they depend on the chosen design for the $x$’s. The max value is at the center of the design.

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 5$</td>
<td>0.63</td>
<td>0.90</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>0.81</td>
<td>0.95</td>
</tr>
<tr>
<td>$n = 20$</td>
<td>0.90</td>
<td>0.97</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0.98</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Table 5.3: Extreme values of $\sqrt{1 - h_i}$ for the fixed design.

5.4 Bootstrapping the Pairs

• Often, the values of $X$ are not fixed in advance but are rather resulting from random sampling the observed units in a population. So we observe the pairs $(x_i, y_i)$ and obtain $\mathcal{X}$.

– The sampling theory above is still valid if we add an “exogeneity condition” on the $X$’s, namely:

$$\forall i, \ E(x_i' \epsilon_i) = 0$$

– We need also an additional regularity assumptions on $X$ to get the asymptotics:

$$\forall i, \ E(x_i' x_i) = V > 0$$

– Bootstrapping on the residuals is OK, if we work conditionnally on the observed values $X_1, \ldots, X_n$. 
• How to take into account for the randomness of the $X$’s?

  – The idea then is to bootstrap the pairs by sampling with replacement in the set $\mathcal{X}$

  – This produce a pseudo-sample $\mathcal{X}^*$:

  $\mathcal{X}^* = \{(x^*_i, y^*_i) \mid i = 1, \ldots, n\}$

  – For each pseudo-sample we compute

  $$\hat{\beta}^* = (X'^*X^*)^{-1}X'^*y^*,$$

  where $X^*$ is the design matrix ($n \times p$):

  $$X^* = \begin{pmatrix}
  x^*_1 \\
  x^*_2 \\
  \vdots \\
  x^*_n
  \end{pmatrix}$$

  where, remember, $x^*_i$ has a one in first position if the model has an intercept.

  – Here we have

  $$\hat{\sigma}^* = \sqrt{\frac{1}{n-p}(y^* - X^*\hat{\beta}^*)'(y^* - X^*\hat{\beta}^*)}$$

  – The bootstrap analog of the root for $\beta_j$ is then:

  $$W^*_j = \frac{\hat{\beta}^*_j - \hat{\beta}_j}{\hat{\sigma}^* \sqrt{c^*_{jj}}},$$

  where $c^*_{jj}$ is the $(j, j)$th element of the matrix $(X'^*X^*)^{-1}$. 
- The bootstrap-\(t\) confidence interval for \(\beta_j\) is:

\[
\left[ \hat{\beta}_j - \hat{\sigma} \sqrt{c_{jj}} w^*_1 - \alpha/2, \hat{\beta}_j - \hat{\sigma} \sqrt{c_{jj}} w^*_\alpha /2 \right]
\]

where \(w^*_a\) is the \(a\)–quantile of Dist\(_{W_j}^*(x)\).

- Advantages and drawbacks:

- If the linear homoskedastic model is true, bootstrapping the residuals will be more accurate and more stable because it follows more the real data generating process. It uses more information.

- Similarly, if we know the parametric family of distributions to which \(F(0, \sigma^2)\) belongs, the parametric bootstrap will be more accurate.

- Bootstrapping the pairs should be more robust to small deviations from the linear homoskedastic model.

  - Bootstrapping the pairs is less sensitive to the model assumptions: we generate data without taking the model into account.

- In case of heteroskedasticity, \(E_F(\epsilon_i^2) = \sigma_i^2\).

  - Bootstrap still consistent but bootstrapping on the pairs should be more stable.

  - Only percentile methods should work here (Bootstrap-\(t\) inappropriate root).

  - We can also use the Wild bootstrap (Mammen, 1993).
• **Example for an homoskedastic case**

- Same scenario as above except that now $X$’s are random variables, we chose the random design $X \sim U(0, 1)$ to be comparable with the above scenario.

- We do not analyze the case $n = 5$: boostrapping on the pairs implies drawing 5 rows among 5 original rows, this produces too many singular $X^*X^*$ in the full Monte-Carlo experiment (5 identical rows with probability $1/625$ at each bootstrap draw for each Monte-Carlo trial).

- Table 5.4 show the results when bootstrapping on the residuals and Table 5.5, when bootstrapping on the the pairs: no significant differences.

  - First row of two tables should be the same (2 different MC experiments).

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 10$</th>
<th>$n = 20$</th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
</tr>
<tr>
<td>As. Normal</td>
<td>4.441</td>
<td>0.923</td>
<td>3.085</td>
<td>0.934</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>3.976</td>
<td>0.889</td>
<td>2.933</td>
<td>0.918</td>
</tr>
<tr>
<td>Bootstrap-t</td>
<td>5.188</td>
<td>0.952</td>
<td>3.281</td>
<td>0.951</td>
</tr>
</tbody>
</table>

Table 5.4: *Performances of confidence intervals for $\beta_2 = 1$, Bootstrapping on the residual. Same scenario as in Table 5.1 but here, $X \sim U(0,1)$, over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$. 
Table 5.5: Performances of confidence intervals for $\beta_2 = 1$, Bootstrapping on the pairs. Same scenario as in Table 5.1 but here, $X \sim U(0,1)$, over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.

- Example of heteroscedastic case

- Scenario:

$$y = 1 + 2x + \epsilon$$

where $\sigma(x) = x$ and $x \sim U(0,20)$.

- Table 5.6 show the results when bootstrapping on the residuals and Table 5.7 on the pairs: significant differences.

- Results

  - First row of two tables should be the same (2 different MC experiments)
  - Asymptotic is inappropriate (even with $n = 500$)
  - As expected, bootstrapping on the pairs outperforms the other bootstrap (in particular when $n \geq 100$).

Table 5.6: Performances of confidence intervals for $\beta_2 = 2$, Bootstrapping on the residual, over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$. 

<table>
<thead>
<tr>
<th>sample size</th>
<th>$n = 10$</th>
<th>$n = 20$</th>
<th>$n = 100$</th>
<th>$n = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>method</td>
<td>length</td>
<td>coverage</td>
<td>length</td>
<td>coverage</td>
</tr>
<tr>
<td>As. Normal</td>
<td>4.425</td>
<td>0.917</td>
<td>3.053</td>
<td>0.936</td>
</tr>
<tr>
<td>Perc. Boot.</td>
<td>4.725</td>
<td>0.911</td>
<td>3.011</td>
<td>0.922</td>
</tr>
<tr>
<td>Bootstrap-t</td>
<td>6.273</td>
<td>0.951</td>
<td>3.356</td>
<td>0.926</td>
</tr>
</tbody>
</table>
Table 5.7: Performances of confidence intervals for $\beta_2 = 2$, Bootstrapping on the pairs, over another $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.

5.5 The Wild Bootstrap

- Proposed by Beran (1986), Liu (1988) and Mammen (1993) for heteroskedasticity of unknown form. See also Davidson and Flachaire (2001). The model

$$y_i = x_i\beta + u_i$$

where $\epsilon_i$ are independent with $E(u_i) = 0$ and $E(u_i^2) = \sigma_i^2$. Write $\Omega = Cov(u)$.

- MacKinnon and White (1985) suggest different forms of HCCME $\hat{\Omega}$ (heteroskedastic consistent covariance matrix estimator) of $\Omega$. Basic estimator of $Var(\hat{\beta}_{OLS})$ is given by

$$Var(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1},$$

where typically $\hat{\Omega} = \text{diag}(\hat{u}_1^2, \ldots, \hat{u}_n^2)$ and $\hat{u}_i$ are the (centered) OLS residuals (refinement: $\hat{u}_i$ is replaced by $\hat{u}_i/\sqrt{1-h_i}$).

- The bootstrap DGP:

$$y_i^* = x_i\hat{\beta} + u_i^*,$$

where $u_i^* = \hat{u}_i\epsilon_i$ where $\epsilon_i$ are mutually independent drawings completely independent of the original data and such that $E(\epsilon_i) = 0$, $E(\epsilon_i^2) = 1$ and $E(\epsilon_i^3) = 1$. Mammen (1993) suggests the use of a
two-point distribution for $\varepsilon_i$:

$$\varepsilon_i = \begin{cases} 
(1 - \sqrt{5})/2 & \text{with probability } p = (5 + \sqrt{5})/10 \\
(1 + \sqrt{5})/2 & \text{with probability } 1 - p 
\end{cases}$$

- NB: The roots are here:

$$W_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{cov}_{jj}}}, \quad \text{and} \quad W^*_j = \frac{\hat{\beta}^*_j - \hat{\beta}_j}{\sqrt{\text{cov}^*_{jj}}}$$

where $\text{cov}_{jj}$ is the $(j, j)$th element of the matrix $(X'X)^{-1}X\hat{\Omega}X(X'X)^{-1}$ and $\text{cov}^*_{jj}$ is the $(j, j)$th element of the matrix $(X'X)^{-1}X\hat{\Omega}^*X(X'X)^{-1}$.

- **More Monte-Carlo Examples with Wild Bootstrap**

**Homoskedastic case:** simulated data according the model:

- The regression model is

$$y_i = 2 + x_i + \varepsilon_i$$

where $\varepsilon_i \sim \Gamma(a, c) - \mu$, with $\mu = ac$, $a = 2$, $c = 1/\sqrt{2}$, so $\sigma_\varepsilon = 1$.

- $X$’s are random variables $X \sim U(1, 20)$.

- We do not analyze the case $n = 5$ for bootstrapping on the pairs (drawing 5 rows among 5 original rows produces too many singular $X^*X^*$).

- Table 5.8 show the results
<table>
<thead>
<tr>
<th>method</th>
<th>length</th>
<th>coverage</th>
<th>length</th>
<th>coverage</th>
<th>length</th>
<th>coverage</th>
<th>length</th>
<th>coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS As. N.</td>
<td>0.372</td>
<td>0.838</td>
<td>0.235</td>
<td>0.915</td>
<td>0.163</td>
<td>0.934</td>
<td>0.072</td>
<td>0.946</td>
</tr>
<tr>
<td>Perc. (Res)</td>
<td>0.284</td>
<td>0.753</td>
<td>0.210</td>
<td>0.885</td>
<td>0.155</td>
<td>0.922</td>
<td>0.071</td>
<td>0.944</td>
</tr>
<tr>
<td>Boot-t (Res)</td>
<td>0.627</td>
<td>0.937</td>
<td>0.275</td>
<td>0.954</td>
<td>0.174</td>
<td>0.945</td>
<td>0.073</td>
<td>0.950</td>
</tr>
<tr>
<td>Perc. (Pairs)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>0.255</td>
<td>0.903</td>
<td>0.160</td>
<td>0.919</td>
<td>0.071</td>
<td>0.943</td>
</tr>
<tr>
<td>Boot-t (Pairs)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>0.333</td>
<td>0.946</td>
<td>0.178</td>
<td>0.943</td>
<td>0.073</td>
<td>0.948</td>
</tr>
<tr>
<td>HCCME As. N.</td>
<td>0.229</td>
<td>0.684</td>
<td>0.190</td>
<td>0.849</td>
<td>0.146</td>
<td>0.907</td>
<td>0.071</td>
<td>0.944</td>
</tr>
<tr>
<td>Perc. (Wild)</td>
<td>0.225</td>
<td>0.671</td>
<td>0.187</td>
<td>0.834</td>
<td>0.145</td>
<td>0.895</td>
<td>0.070</td>
<td>0.938</td>
</tr>
<tr>
<td>Boot-t (Wild)</td>
<td>0.533</td>
<td>0.828</td>
<td>0.251</td>
<td>0.887</td>
<td>0.160</td>
<td>0.921</td>
<td>0.072</td>
<td>0.941</td>
</tr>
</tbody>
</table>

Table 5.8: HOMOSKEDASTIC SCENARIO. Performances of confidence intervals for $\beta_2 = 1$, Bootstrapping on the residual, on the pairs and Wild bootstrap, over $M = 1000$ Monte-Carlo experiments with $B = 2000$.

Heteroscedastic case: simulated data according

- The regression model is

  $$ y_i = 2 + x_i + \epsilon_i $$

  where $\epsilon_i \sim |x_i - 10| \times (\Gamma(a, c) - \mu)$, with $\mu = ac$, $a = 2$, $c = 1/\sqrt{2}$, so $\sigma_\epsilon = 1$.

- $X$’s are random variables $X \sim U(1, 20)$.

- Table 5.9 show the results.

- NB: Wild bootstrap allows test of hypothesis in the presence of heteroskedasticity (see below).
Table 5.9: HETEROSKEDASTIC SCENARIO. Performances of confidence intervals for $\beta_2 = 1$, Bootstrapping on the residual, on the pairs and Wild bootstrap, over $M = 1000$ Monte-Carlo experiments with $B = 2000$.

5.6 Testing Hypothesis

- Suppose we want to test the hypothesis $H_0 : \beta_j = 0$ against the alternative $H_1 : \beta_j > 0$

  - The test statistics can be
    \[ T(\mathbf{x}) = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{c_{jj}}} \]
    where $c_{jj}$ is the $(j,j)$th element of the matrix $(X'X)^{-1}$.

  - The $p$-value is given by:
    \[ p = P(T(\mathbf{x}) \geq T_{observed} \mid H_0) \]

  - We have to estimate $p$ by the bootstrap algorithm. How to generate the samples under $H_0$?
    - We estimate $\beta$ under the null hypothesis:
      \[ \hat{\beta}^{(0)} = (\tilde{X}'\tilde{X})^{-1} \tilde{X}'y \]
where $\tilde{X}$ is the design matrix $X$ with the $j$th column deleted.

- We define the residuals under $H_0$ as

$$\hat{e}^{(0)} = y - \tilde{X}\hat{\beta}^{(0)}$$

- We draw with replacement $n$ values from $\hat{e}_i^{(0)}, i = 1, \ldots, n$ to produce the bootstrap residuals $\hat{e}_i^{(0)*}, i = 1, \ldots, n$.

- The bootstrap sample under $H_0$ is $\mathcal{X}^{(0)*} = \{(x_i, y_i^*) | i = 1, \ldots, n\}$ where

$$y_i^* = \tilde{x}_i \hat{\beta}^{(0)} + \hat{e}_i^{(0)*}.$$  

- Finally we have the bootstrap estimation of $p$

$$\hat{p} = P(T(\mathcal{X}^*) \geq T_{observed} | \mathcal{X}, H_0)$$

where

$$T(\mathcal{X}^*) = \frac{\hat{\beta}_j^*}{\hat{\sigma}^* \sqrt{c_{jj}}}$$

where $\hat{\beta}_j^*$ is the $j$th element of $\hat{\beta}^* = (X'X)^{-1}X'y^*$ and $\hat{\sigma}^2 = (y^* - X\hat{\beta}^*)(y^* - X\hat{\beta}^*)/(n - p)$.

- Testing a linear restrictions $H_0 : A\beta = a$ where $A : (q \times p)$ and $a : (q \times 1)$.

- Here, the OLS under the null is (see Härdle and Simar, 2003, p.196):

$$\hat{\beta}^{(0)} = \hat{\beta} - (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - a)$$
A test statistics would be

\[ T(\mathcal{X}) = \frac{||y - X\hat{\beta}^{(0)}||^2}{||y - X\hat{\beta}||^2} - 1 \geq 0 \]

and the \( p \)-value is

\[ \hat{p} = P(T(\mathcal{X}^*) \geq T_{observed} | \mathcal{X}, H_0) \]

– same approach as above to generate \( \mathcal{X}^* \) by using \( \hat{\beta}^{(0)} \).

– NB: if \( \epsilon \sim N(0, \sigma^2) \), we have an exact test:

\[ \frac{n-p}{q} T(\mathcal{X}) \sim F_{q,n-p} \text{ if } H_0 \text{ is true} \]

• In case of heteroskedasticity, the Wild bootstrap has to be used.
• Example

- We have the following bivariate data \((n = 8)\):

\[
\begin{align*}
  x &= (0.6093 \ 0.9303 \ 0.3414 \ 0.1874 \ 0.0160 \ 0.6577 \ 0.2924 \ 0.0758) \\
  y &= (2.0663 \ 2.4426 \ 1.9896 \ 2.7825 \ 1.6037 \ 2.5204 \ 4.5603 \ 1.0824)
\end{align*}
\]

The data are presented in Figure 5.1 with the least squares fit.

![Scatterplot of the data and OLS fit.](image)

- The OLS fit gives \(\hat{\beta} = (2.1458 \ 0.6048)'\), we test \(H_0 : \beta_2 = 0\) against \(H_1 : \beta_2 > 0\).

  - Under \(H_0 : \beta_2 = 0\) we have \(\tilde{\beta}_1 = \bar{y} = 2.3810\)

  - The confidence intervals for \(\beta_2\) at 95% are \((B = 5000)\)
    
    Asymptotic Normal: \([-1.9590, 3.1686]\)
    
    Percentile Bootstrap: \([-1.5177, 3.0111]\)
    
    Bootstrap-t method: \([-2.7733, 3.4138]\)

- The \(p\)-value are computed as
  
  Student case \(p = 0.3188\)
  
  Bootstrap \(\hat{p} = 0.3079\)

- There is no sample evidence to reject \(H_0\) at the level 5%.
5.7 Bootstrapped Prediction Intervals

Consider the problem of the prediction of a new value $y_f$ independent from the $y_i$, $i = 1, \ldots, n$ for a given level of the vector $x_f$ (Stine, 1985).

- Consider first the real world.

  - $y_f$ is generated by $\mathcal{P}$ as follows:
    \[ y_f = x_f \beta + \epsilon_f, \text{ where } \epsilon_f \sim F(0, \sigma^2). \]

  - The prediction is given by
    \[ \hat{y}_f = x_f \hat{\beta} \]
    where $\hat{\beta}$ is the estimator of $\beta$ based on the sample $\mathcal{X}$ of size $n$ generated from $\mathcal{P}$.

  - The error of prediction is
    \[ e_f = y_f - \hat{y}_f = x_f(\beta - \hat{\beta}) + \epsilon_f \]
    we have:
    \[ E_F(e_f) = 0 \]
    \[ Var_F(e_f) = \sigma^2 \left( 1 + x_f(X'X)^{-1}x_f' \right) = \sigma^2(1 + h_f) \]

  - The natural root is
    \[ R = \frac{y_f - \hat{y}_f}{\hat{\sigma} \sqrt{1 + h_f}} \]
– Under Normal assumption, we have $R \sim N(0, 1)$. The normal prediction interval is:
\[
\left[ \hat{y}_f - z_{1-\alpha/2} \hat{\sigma} \sqrt{1 + h_f}, \hat{y}_f - z_{\alpha/2} \hat{\sigma} \sqrt{1 + h_f} \right]
\]
where $z_a$ is the $a$-quantile of the $N(0, 1)$.

• Consider now the bootstrap world, where $\hat{P}$ is characterized by:
\[
y = x\hat{\beta} + \epsilon^* \quad \text{where } \epsilon^* \sim F_n
\]
– $y^*_f$ is generated by $\hat{P}$ as follows:
\[
y^*_f = x_f\hat{\beta} + \epsilon^*_f \quad \text{where } \epsilon^*_f \sim F_n.
\]
– The prediction is given by
\[
\hat{y}^*_f = x_f\hat{\beta}^*
\]
where $\hat{\beta}^*$ is the estimator of $\hat{\beta}$ based on the sample $X^*$ of size $n$ generated from $\hat{P}$.

– The error of prediction is
\[
\epsilon^*_f = y^*_f - \hat{y}^*_f = x_f(\hat{\beta} - \hat{\beta}^*) + \epsilon^*_f
\]

– The bootstrap root is
\[
R^* = \frac{y^*_f - \hat{y}^*_f}{\hat{\sigma}^* \sqrt{1 + h_f}},
\]
where $\hat{\sigma}^*_2$ is computed from the sample $X^*$.
- The bootstrap-$t$ prediction interval is:

$$\left[ \hat{y}_f - u_{1-\alpha/2} \hat{\sigma} \sqrt{1 + h_f}, \hat{y}_f - u_{\alpha/2} \hat{\sigma} \sqrt{1 + h_f} \right]$$

where $u_a^*$ is the $a$-quantile of $\text{Dist}_{R}^*$.

- The algorithm can be summarized as follows:

1. Compute $\hat{\beta} = (X'X)^{-1}X'y$, the OLS residuals $\hat{\epsilon} = y - X\hat{\beta}$, $\hat{\sigma}^2 = \hat{\epsilon}'\hat{\epsilon}/(n - p)$ and the prediction $\hat{y}_f = x_f\hat{\beta}$.
2. From the bootstrap sample $X^*$ compute, in the same way, $\hat{\beta}^*$, $\hat{\sigma}^*$ and the prediction in the bootstrap world $\hat{y}_f^* = x_f\hat{\beta}^*$.
3. Generate the value of $y_f$ in the bootstrap world: $y_f^* = x_f\hat{\beta} + \epsilon_f^*$, where $\epsilon_f^*$ is one drawn from the $n$ OLS residuals $\hat{\epsilon}$.
4. Compute the root

$$R^* = \frac{y_f^* - \hat{y}_f^*}{\hat{\sigma}^* \sqrt{1 + h_f}}.$$ 

5. Redo steps [2] to [4], $B$ times to derive the empirical distribution of $R^*$: $\text{Dist}_{R}^*$, from which we obtain the appropriate quantiles $u_a^*$. 
Example

- Back to the preceding example (see Figure 5.1).
  - The data:
    \[ x = (0.6093, 0.9303, 0.3414, 0.1874, 0.0160, 0.6577, 0.2924, 0.0758) \]
    \[ y = (2.0663, 2.4426, 1.9896, 2.7825, 1.6037, 2.5204, 4.5603, 1.0824) \]
  - The OLS fit gives \( \hat{\beta} = (2.1458, 0.6048)' \) and \( \hat{\sigma} = 1.096 \)
  - Prediction at \( x_f = 0.5 \).
    - We obtain \( \hat{y}_f = 2.4482 \)
    - The prediction intervals (\( B = 5000 \)):
      - Asymptotic Normal: \([0.1501, 4.7463]\)
      - Nonparametric Bootstrap: \([-3.3438, 4.4533]\)
  - Remark: This sample was obtained by simulation \( n = 8 \) values of \( X \sim U(0, 1) \) and \( \epsilon \sim \text{Shifted - Gamma} \) with mean zero and variance 1. The model is \( y = 2 + x + \epsilon \).
  - So, the normal approximation is not appropriate here!
Chapter 6

Iterated Bootstrap

6.1 Estimating the variance for prepivoting

- If we have an asymptotical pivotal root

\[ S = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\sigma(F)} \right) \sim AN(0, 1), \]

where \( \sigma^2(F) = n \text{Var}_F(T(\mathcal{X})) \) and if \( \hat{\sigma}(\mathcal{X}) \) is available:

- we could use the studentized root for the bootstrap:

\[ U = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\hat{\sigma}(\mathcal{X})} \right) \sim AN(0, 1), \]

where \( \hat{\sigma}(\mathcal{X}) \) is a consistent estimator of \( \sigma(F) \).

- The bootstrap-t confidence interval for \( \theta(F) \) is:

\[
\left[ T(\mathcal{X}) - u^*(1 - \frac{\alpha}{2}) \frac{\hat{\sigma}}{\sqrt{n}}, T(\mathcal{X}) - u^*(\frac{\alpha}{2}) \frac{\hat{\sigma}}{\sqrt{n}} \right]
\]

where \( u^*(a) \) is the \( a \)-quantile of \( Dist_{U^*}(x) \) the bootstrap distribution of

\[ U^* = \sqrt{n} \left( \frac{T(\mathcal{X}^*) - \theta(F_n)}{\hat{\sigma}(\mathcal{X}^*)} \right) \]
• Suppose that $\hat{\sigma}(\mathcal{X})$ is not available.

- The idea: estimate $\sigma^2(F)$ by a bootstrap method to get $\hat{\sigma}^2$

$$\hat{\sigma}^2 = n \text{Var}_{F_n}(T(\mathcal{X}^*))$$

- Therefore, the original root in the "real world" becomes:

$$W = \sqrt{n} \left( \frac{T(\mathcal{X}) - \theta(F)}{\hat{\sigma}^*} \right)$$

where $\hat{\sigma}^*$ is thus obtained by a first level bootstrap.

- The confidence interval for $\theta(F)$ is then:

$$\left[ T(\mathcal{X}) - w(1 - \frac{\alpha}{2}) \frac{\hat{\sigma}^*}{\sqrt{n}}, T(\mathcal{X}) - w(\frac{\alpha}{2}) \frac{\hat{\sigma}^*}{\sqrt{n}} \right]$$

where $w(a)$ is the $a-$quantile of $\text{Dist}_W(x)$ and $\hat{\sigma}^*$ is the first level bootstrap estimate of $\sigma$.

- Since $\text{Dist}_W(x)$ is unknown, we will use a second level bootstrap to approximate it.

- The bootstrap analog of $W$ in the "bootstrap world" becomes $W^*$:

$$W^* = \sqrt{n} \left( \frac{T(\mathcal{X}^*) - \theta(F_n)}{\hat{\sigma}^{**}} \right)$$

- where $(\hat{\sigma}^{**})^2 = n \text{Var}_{F_n^*}(T(\mathcal{X}^{**}))$ is a bootstrap estimator in the bootstrap world of $\hat{\sigma}^2 = n \text{Var}_{F_n}(T(\mathcal{X}^*))$.

* $F_n^*$ is the empirical cdf of the sample $\mathcal{X}^*$ and $\mathcal{X}^{**}$ is sample drawn from $F_n^*$ (drawn from $\mathcal{X}^*$ with replacement).

* $\hat{\sigma}^{**}$ has to be obtained for each bootstrap sample $\mathcal{X}^*$.
- \( \hat{\sigma}^{**} \) has to be derived through a **second level** bootstrap: a bootstrap nested inside the first level bootstrap world.

- This is called the **double bootstrap**

- The final bootstrap-t confidence interval for \( \theta(F) \) is

\[
\left[ T(\mathcal{X}) - w^*(1 - \frac{\alpha}{2}) \frac{\hat{\sigma}^*}{\sqrt{n}}, T(\mathcal{X}) - w^*(\frac{\alpha}{2}) \frac{\hat{\sigma}^*}{\sqrt{n}} \right]
\]

where \( w^*(a) \) is the \( a \)-quantile of \( Dist_W^*(x) \) and \( \hat{\sigma}^* \) is the **first level bootstrap** estimate of \( \sigma \).

- Generally, the double bootstrap recovers the \( O_p(n^{-1}) \) precision: second order accuracy, as the bootstrap-t.

**The algorithm**

[1 ] For each first level bootstrap iteration \( b_1 \), draw a sample \( \mathcal{X}^{*(b_1)} \) from \( \mathcal{X} \): this provides \( T^{*(b_1)} = T(\mathcal{X}^{*(b_1)}) \).

[2 ] Estimate the variance of \( T(\mathcal{X}^{*(b_1)}) \Rightarrow \) second level bootstrap:

[2.1 ] Draw \( B_2 \) samples \( \mathcal{X}^{**(b_2)} \) from \( \mathcal{X}^{*(b_1)} \), for \( b_2 = 1, \ldots, B_2 \).

[2.2 ] Compute

\[
Var^{**}(T^{*(b_1)}) \approx \frac{1}{B_2} \sum_{b_2=1}^{B_2} T^2(\mathcal{X}^{**(b_2)}) - \left( \frac{1}{B_2} \sum_{b_2=1}^{B_2} T(\mathcal{X}^{**(b_2)}) \right)^2
\]

[2.3 ] Then:

\[
\hat{\sigma}^{**(b_1)} = \sqrt{n} \left( Var^{**}(T^{*(b_1)}) \right)^{1/2}
\]
[3] Compute the bootstrap value of the root:

\[ W^{*}(b_1) = \sqrt{n} \left( \frac{T(X^{*}(b_1)) - \theta(F_n)}{\hat{\sigma}^{**}(b_1)} \right) \]

[4] Redo the steps [1]–[3] for \( b_1 = 1, \ldots, B_1 \). From the empirical cdf of \( W^{*(b_1)}, b_1 = 1, \ldots, B_1 \) we have \( Dist^{*}_{W}(x) \).

[5] Compute also \( \hat{\sigma}^{*} \), the first level bootstrap estimate of \( \sigma \):

\[ \hat{\sigma}^{*} = \sqrt{n} \left( \frac{1}{B_1} \sum_{b_1=1}^{B_1} T^2(X^{*(b_1)}) - \left( \frac{1}{B_1} \sum_{b_1=1}^{B_1} T(X^{*(b_1)}) \right)^2 \right)^{1/2} \]

- The bootstrap-t confidence interval for \( \theta(F) \) is:

\[ \left[ T(X) - w^{*}(1 - \frac{\alpha}{2}) \frac{\hat{\sigma}^{*}}{\sqrt{n}}, T(X) - w^{*}(\frac{\alpha}{2}) \frac{\hat{\sigma}^{*}}{\sqrt{n}} \right] \]

where \( w^{*}(a) \) is the \( a \)-quantile of \( Dist^{*}_{W}(x) \).

- Remark: Computer intensive method: \( B_1 \times B_2 \) Monte-Carlo loops where both \( B_1 \) should be, say \( \geq 2000 \) and \( B_2 \) should be, say \( \geq 200 \).
**Example:** Let’s come back to the duration data $n = 10$:

$$\mathcal{X} = (X_1, \ldots, X_{10}) = (1, 5, 12, 15, 20, 26, 78, 145, 158, 358)$$

We have $\bar{x} = 81.80$ and $s = 112.94$.

- Here we can use the bootstrap-$t$ (see above Table 2.1), because we have $s$ estimator of $\sigma$.
- Suppose instead, we use the double bootstrap to estimate $\sigma$. The root is:

$$W = \sqrt{n} \left( \frac{\bar{X} - \mu}{\hat{\sigma}^*} \right)$$

The Double-bootstrap-$t$ confidence interval is given by

$$\left[ \bar{X} - w^*(1 - \frac{\alpha}{2}) \frac{\hat{\sigma}^*}{\sqrt{n}}, \bar{X} - w^*(\frac{\alpha}{2}) \frac{\hat{\sigma}^*}{\sqrt{n}} \right]$$

where $w^*(a)$ is the $a$-quantile of the empirical bootstrap distribution of

$$W^* = \sqrt{n} \left( \frac{\bar{X}^* - \bar{X}}{\hat{\sigma}^{**}} \right)$$

- Note that here $s = 112.9383$ and $\hat{\sigma}^* = 106.8983$ (with $B1 = 2000$).
- The results of Table 6.1 compare the two bootstrap-$t$ approaches.
- Figure 6.1 shows the empirical bootstrap densities of the two roots (usual studentized and bootstrap studentized).
<table>
<thead>
<tr>
<th>method</th>
<th>lower limit</th>
<th>upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usual bootstrap-t</td>
<td>25.1980</td>
<td>300.7924</td>
</tr>
<tr>
<td>Double bootstrap-t</td>
<td>24.0624</td>
<td>288.3107</td>
</tr>
</tbody>
</table>

Table 6.1: Confidence intervals, duration data with $n = 10$. Here $B_1 = 2000$ and $B_2 = 200$.

Figure 6.1: Empirical bootstrap densities of the two roots: usual studentized (solid) and bootstrap studentized (dash-dotted).

- Part of the Matlab code for the double bootstrap:

```matlab
NParXbar=[];
Studboot=[];
Doublboot=[];
for b1=1:B1
    xb=boot(x);
    mb=mean(xb);
    sb=std(xb);
    % second level bootstrap for estimating the std of mb
    mbb=[];
    for b2=1:B2
        xbb=boot(xb);
        mbb=[mbb;mean(xbb)];
    end
    sbb=sqrt(n)*std(mbb,1);
    %
    studb=sqrt(n)*(mb-m)/sb;
    studbb=sqrt(n)*(mb-m)/sbb;
    %
    NParXbar=[NParXbar;mb];
    Studboot=[Studboot;studb];% provide the Bootstrap-t sampl. distr.
    Doublboot=[Doublboot;studbb];% provide the Double-bootstrap sampl. distr.
end
sbstar=std(NParXbar,1);
```
6.2 Calibrating coverage errors of CI

- We know the accuracy of the percentile method and/or of the basic bootstrap method is only of the first order.
  - The coverage probability of the resulting confidence intervals could be different from the desired nominal level.
  - The idea: use a second level bootstrap to estimate the coverage probability of the obtained confidence interval and recalibrate the interval to improve the coverage.

- **First level bootstrap**
  - The root and its bootstrap analog are
    \[
    W = T(\mathcal{X}) - \theta(F) \\
    W^* = T(\mathcal{X}^*) - \theta(F_n)
    \]
    where very often \( \theta(F_n) = T(\mathcal{X}) \).
  - The basic bootstrap method provides a confidence interval for \( \theta(F) \):
    \[
    I_\alpha = [T(\mathcal{X}) - w^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) - w^*(\frac{\alpha}{2})]
    \]
  - Here \( w^*(a) \) is the \( a \)-quantile of \( Dist^*_W(x) \), the empirical cdf of \( W^* \) obtained through the \( B_1 \) Monte-Carlo replications of the first level bootstrap.
The true coverage probability is $\pi(\alpha)$, and we hope that

$$\pi(\alpha) = P_F(\theta(F) \in \mathcal{I}_\alpha) \approx 1 - \alpha$$

For the basic bootstrap (and the percentile) the order of the error is usually $O_p(n^{-\frac{1}{2}})$.

Second level bootstrap to estimate $\pi(\alpha)$ and then to recalibrate $\mathcal{I}_\alpha$ in order to improve the accuracy and achieve the order $O_p(n^{-1})$.

**Second level bootstrap**

The second level bootstrap pseudo-sample is

$$\mathcal{X}^{**} = (X^{**}_1, \ldots, X^{**}_n)$$

where $X^{**}_i$ is drawn with replacement from $\mathcal{X}^* = (X^*_1, \ldots, X^*_n)$ for $i = 1, \ldots, n$.

Here we have $X^{**}_i \sim F^*_n$ where $F^*_n$ is the empirical cdf of $\mathcal{X}^* = (X^*_1, \ldots, X^*_n)$.

In the bootstrap world, the second level bootstrap root is

$$W^{**} = T(\mathcal{X}^{**}) - \theta(F^*_n)$$

where, very often $\theta(F^*_n) = T(\mathcal{X}^*)$.

The same method as in the first level above, provides a second level bootstrap confidence interval for $\theta(F_n)$ (which is known!):
\[ I^*_\alpha = [T(X^*) - w^{**}(1 - \frac{\alpha}{2}), T(X^*) - w^{**}(\frac{\alpha}{2})] \]

- Here \( w^{**}(a) \) is the \( a \)-quantile of \( \text{Dist}^{**}_W(x) \), the empirical cdf of \( W^{**} \) obtained through the \( B_2 \) Monte-Carlo replications of the second level bootstrap.

- So, at each iteration \( b_1, b_1 = 1, \ldots, B_1 \) of the first level bootstrap, we obtain:

\[ I^*_\alpha = I^{*(b_1)}_{\alpha} \]

- Since \( \theta(F_n) \) is known, we can estimate \( \pi(\alpha) \) by the observed proportion of time \( \theta(F_n) \in I^*_\alpha \):

\[ \hat{\pi}(\alpha) = P_{F_n}(\theta(F_n) \in I^*_\alpha) \]

- This is obtained through the following:

\[ \hat{\pi}(\alpha) \approx \frac{1}{B_1} \sum_{b_1=1}^{B_1} I(\theta(F_n) \in I^{*(b_1)}_{\alpha}) \]

- This is the proportion of times that \( I^*_\alpha \) covers \( \theta(F_n) \) in repeating both levels of the Bootstrap many times.
Calibration

- Solve in $\alpha$
  \[ \hat{\pi}(\alpha) = 1 - \alpha_0 \]
  where $1 - \alpha_0$ is the desired nominal coverage probability (say, 0.95).
- Let $\hat{\alpha}$ be the solution:
  \[ \hat{\pi}(\hat{\alpha}) = 1 - \alpha_0 \]
- The calibrated confidence interval is $I_{\hat{\alpha}}$:
  \[ I_{\hat{\alpha}} = [T(X) - w^*(1 - \frac{\hat{\alpha}}{2}), T(X) - w^*(\frac{\hat{\alpha}}{2})] \]
  where $w^*(a)$ is, as above, the $a-$quantile of $Dist^*_W(x)$
- The true coverage probability is $\pi(\hat{\alpha})$
  \[ \pi(\hat{\alpha}) = P_{F} (\theta(F) \in I_{\hat{\alpha}}) \approx 1 - \alpha_0 \]
  The order of the error is now generally $O_p(n^{-1})$.

The algorithm

[1 ] At each first level bootstrap iteration $b_1$, draw a sample $X^{*(b_1)}$ from $X$: this provides $T^{*(b_1)} = T(X^{*(b_1)})$. Denote $F_n^{*(b_1)}$ the empirical cdf of the sample $X^{*(b_1)}$.

[2 ] Estimate a confidence interval $I_{\alpha}^{*(b_1)}$ for $\theta(F_n)$ from this estimator $T(X^{*(b_1)})$ ⇒ second level bootstrap:
- [2.1] Draw a sample \( \mathcal{X}^{**(b_2)} \) from \( \mathcal{X}^{*(b_1)} \), compute \( T(\mathcal{X}^{**(b_2)}) \), so:

\[
W^{**(b_1)} = T(\mathcal{X}^{**(b_2)}) - \theta(F_n^{*(b_1)})
\]

- [2.2] Redo the step [2.1] for \( b_2 = 1, \ldots, B_2 \) and form the empirical cdf of \( W^{**(b_1)} \) denoted by \( Dist_{W}^{**(b_1)}(x) \).

[3] Compute the quantiles \( w^{**(b_1)}(\alpha/2) \) and \( w^{**(b_1)}(1 - \alpha/2) \) from \( Dist_{W}^{**(b_1)}(x) \).

[4] The second level bootstrap confidence interval for \( \theta(F_n) \) is:

\[
I^{*(b_1)}_{\alpha} = \left[ T(\mathcal{X}^{*(b_1)}) - w^{**(b_1)}(1 - \frac{\alpha}{2}), T(\mathcal{X}^{*(b_1)}) - w^{**(b_1)}(\frac{\alpha}{2}) \right].
\]

[5] Redo the steps [1]–[4] for \( b_1 = 1, \ldots, B_1 \): this provides \( B_1 \) values of \( W^{*(b_1)} = T(\mathcal{X}^{*(b_1)}) - \theta(F_n) \), with their empirical cdf \( Dist_{W}^{*(b_1)}(x) \) and \( B_1 \) values of \( I^{*(b_1)}_{\alpha} \).

[6] Compute

\[
\hat{\pi}(\alpha) = \frac{1}{B_1} \sum_{b_1=1}^{B_1} I(\theta(F_n) \in I^{*(b_1)}_{\alpha})
\]

[7] Solve in \( \alpha \) the equation \( \hat{\pi}(\alpha) = 1 - \alpha_0 \) and call \( \hat{\alpha} \) the solution:

\[
\hat{\pi}(\hat{\alpha}) = 1 - \alpha_0
\]

[8] The corrected confidence interval for \( \theta \) is

\[
I_{\hat{\alpha}} = [T(\mathcal{X}) - w^*(1 - \frac{\hat{\alpha}}{2}), T(\mathcal{X}) - w^*(\frac{\hat{\alpha}}{2})]
\]

where \( w^*(a) \) is, as above, the \( a \)–quantile of \( Dist_{W}^{*}(x) \).
The step [7] can be solved as follows:

- we select a set of values $\alpha_1, \ldots, \alpha_k$ near the desired value $\alpha_0$ and compute in step [6] $\hat{\pi}(\alpha_j)$ for those selected values
- $\hat{\alpha}$ is then founded by linear interpolation
- see example below

Remark : This is a computer intensive method : $B_1 \times B_2$ Monte-Carlo loops where $B_1$ and $B_2$ should be large (say, both $\geq 1000$).

Example:

- We simulate one sample of size $n = 20$ from a $N(10, 2^2)$.
- We obtain $\bar{x} = 10.2322$ and $s = 1.7547$.
- The basic bootstrap method gives the 95% confidence interval for $\mu$: $\mu \in [9.4834, 10.9626]$. Note the length of CI is 1.4792.
- The double bootstrap estimated the coverage probabilities for several nominal levels $1 - \alpha$. The result are in Table 6.2
- Figure 6.2 shows the relation between $1 - \alpha$ and $\hat{\pi}(\alpha)$: this allows to determine, by linear interpolation, $\hat{\alpha}$, such that $\hat{\pi}(\hat{\alpha}) \approx 1 - \alpha_0$ for the desired nominal level $\alpha_0$.
- Here, with $\alpha_0 = 0.05$, the corrected level is 0.9650 so $\hat{\alpha} = 0.0350$ and the corrected 95% confidence interval for $\mu$ turns out to be $\mu \in [9.4170, 11.0094]$. Note the length of CI is 1.5924.
Note that here the exact CI is available (Student): it is given by $\bar{x} \pm t_{19,\alpha/2}s/\sqrt{n}$. In our case here at 95% level we obtain $\mu \in [9.4110, 11.0534]$. Note the length of this CI is 1.6424.

<table>
<thead>
<tr>
<th>$1 - \alpha$</th>
<th>$\hat{\pi}(\alpha)$</th>
<th>confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8000</td>
<td>0.7650</td>
<td>[9.7577, 10.7239]</td>
</tr>
<tr>
<td>0.8500</td>
<td>0.8190</td>
<td>[9.6943, 10.7874]</td>
</tr>
<tr>
<td>0.9000</td>
<td>0.8770</td>
<td>[9.6163, 10.8631]</td>
</tr>
<tr>
<td>0.9100</td>
<td>0.8850</td>
<td>[9.5757, 10.8766]</td>
</tr>
<tr>
<td>0.9200</td>
<td>0.8990</td>
<td>[9.5552, 10.8915]</td>
</tr>
<tr>
<td>0.9300</td>
<td>0.9110</td>
<td>[9.5369, 10.9205]</td>
</tr>
<tr>
<td>0.9400</td>
<td>0.9260</td>
<td>[9.5076, 10.9369]</td>
</tr>
<tr>
<td>0.9500</td>
<td>0.9360</td>
<td>[9.4834, 10.9626]</td>
</tr>
<tr>
<td>0.9600</td>
<td>0.9430</td>
<td>[9.4421, 10.9878]</td>
</tr>
<tr>
<td>0.9700</td>
<td>0.9570</td>
<td>[9.4091, 11.0479]</td>
</tr>
<tr>
<td>0.9800</td>
<td>0.9660</td>
<td>[9.2715, 11.1204]</td>
</tr>
<tr>
<td>0.9900</td>
<td>0.9810</td>
<td>[9.2378, 11.1828]</td>
</tr>
</tbody>
</table>

Table 6.2: Nominal levels and estimated coverages. Here $B_1 = 1000$ and $B_2 = 1000$.

Figure 6.2: Horizontal scale: nominal levels $1 - \alpha$ and vertical scale estimated coverage $\hat{\pi}(\alpha)$. Here, for $\alpha_0 = 0.05$, the calibrated $\alpha$ is $\hat{\alpha} = 1 - 0.9650 = 0.0350$. 
Part of the Matlab code

% set the number of nominal levels wanted
alpha=[0.20 0.15 0.10 0.09 0.08 0.07 0.06 0.05 0.04 0.03 0.02 0.01 ]';
[nci,q]=size(alpha);

mvec=m*ones(nci,1);count=zeros(nci,1);
Wb=[];
for b1=1:B1
    xb=boot(x);mb=mean(xb);
    Wb=[Wb;(mean(xb)-m)];
    % second level bootstrap for estimating the coverage
    Wbb=[];
    for b2=1:B2
        xbb=boot(xb);
        Wbb=[Wbb;(mean(xbb)-mb)];
    end
    mbvec=mb*ones(nci,1);
    temp=[mbvec-prctile(Wbb,100*(1-alpha/2)) mbvec-prctile(Wbb,100*alpha/2)];
    success=(mvec >= temp(:,1) & mvec <= temp(:,2));count=count+success;
end
% estimation of the coverage
pi=count/B1;
disp('nominal alpha and pi(alpha) frequencies')
disp([(1-alpha) pi])
plot(1-alpha,pi,'-.',1-alpha,1-alpha,'-')
axis([0.8 1 0.8 1])

hatalpha=input('Give the value of hatalpha');% here we input hatalpha=0.0350
CI=[m-prctile(Wb,100*(0.975)) m-prctile(Wb,100*0.025)];% original CI
CIcorr=[m-prctile(Wb,100*(1-hatalpha/2)) m-prctile(Wb,100*hatalpha/2)]; % calibrated CI
Chapter 7

The Jacknife

• The jacknife* is a resampling method that predates the bootstrap (Quenouille, 1949 and Tukey, 1958). Two major purposes:
  
  – Reduction of the bias of an estimator
  – Estimation of the variance of an estimator

7.1 The Jacknife method

• Let \( \mathcal{X} = (X_1, \ldots, X_n) \) be a random sample from \( X \) where \( X \sim F \) and let \( \theta(F) \) be the quantity of interest.
  
  – Let \( T_n = T_n(\mathcal{X}) \) be an estimator of \( \theta(F) \) (here we suppose \( T_n \) is a plug-in version of \( \theta \): \( T_n = \theta(F_n) \)).
  
  - Define the leave-one-out estimator of \( \theta(F) \)
    \[
    T_{n-1}^{(i)} = T_{n-1}(\mathcal{X}_{(i)})
    \]
    where \( \mathcal{X}_{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \).

*A preliminary handout from Ingrid Van Keilegom for this section is acknowledged.
- Define the “pseudo-values” $T_{n,i}$

$$T_{n,i} = nT_n - (n - 1)T_{n-1}^{(i)}$$

Note that, in the particular case $T_n(A) = \bar{X}$, we have

$$T_{n,i} = X_i$$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} T_{n,i}.$$  

- The Jacknife estimators:

- Jacknife estimator of the mean of $T_n$:

$$\hat{E}_{jack}(T_n) = \hat{T}_n = \frac{1}{n} \sum_{i=1}^{n} T_{n,i}$$

$$= nT_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^{(i)}$$

This is the average of the $n$ pseudo-values $T_{n,i}$

- Jacknife estimator of the bias of $T_n$:

$$\hat{Bias}_{jack}(T_n) = T_n - \hat{E}_{jack}(T_n)$$

$$= (n - 1) \left[ \frac{1}{n} \sum_{i=1}^{n} T_{n-1}^{(i)} - T_n \right].$$

- Jacknife estimator of the variance of $T_n$:

$$\hat{Var}_{jack}(T_n) = \frac{1}{n} \hat{S}^2_n$$

$$= \frac{1}{n} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (T_{n,i} - \hat{T}_n)^2 \right\}$$

$\hat{S}^2_n$ is the sample variance of the $n$ pseudo-values $T_{n,i}$
• Motivation
  – Suppose that $T_n$ is asymptotically unbiased with
    $$E(T_n) = \theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \ldots$$
  – Now, since $\hat{T}_n = nT_n - \frac{n-1}{n} \sum_{i=1}^{n} T_{n-1}^{(i)}$, we have:
    $$E(\hat{T}_n) = nE(T_n) - (n-1)E(T_{n-1}^{(i)})$$
    $$= n\left(\theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \ldots\right)$$
    $$- (n-1)\left(\theta + \frac{a(\theta)}{n-1} + \frac{b(\theta)}{(n-1)^2} + \ldots\right)$$
    $$= \theta + \frac{b(\theta)}{n} - \frac{b(\theta)}{n-1} + \ldots$$
    $$= \theta - \frac{b(\theta)}{n(n-1)} + \ldots$$
  – $\hat{T}_n$ should be a more accurate estimator of $\theta$

7.1.1 Some examples

• The mean
  – Let $\theta(F) = E(X)$ and $T_n = \bar{X}_n$.
    - We have:
      $$T_{n,i} = n\bar{X}_n - (n-1)\bar{X}_{n-1}^{(i)} = X_i$$
    - So that
      $$\hat{T}_n = \bar{X}_n = T_n$$
      $$\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = S_n^2$$
We obtain

\[ \hat{\text{Bias}}_{\text{jack}}(\bar{X}_n) = 0 \]
\[ \hat{\text{Var}}_{\text{jack}}(\bar{X}_n) = \frac{S_n^2}{n} \]

• **The variance**

  - Here \( \theta(F) = \text{Var}(X) \) and \( T_n = \theta(F_n) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \).

    - Some calculations show that
      \[ T_{n,i} = \frac{n}{n-1} (X_i - \bar{X}_n)^2 \]

    - So that:
      \[ \hat{T}_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = S_n^2 \]

    - Note that \( E(T_n) = \frac{n-1}{n} \theta \) and \( E(\hat{T}_n) = \theta \).

• **Failure of the jackknife**

  The statistics to be analyzed have to be “smooth” in the data.

  - **The Median:**

    - Let \( \theta(F) = F^{-1}(1/2) \) and \( T_n = (X_{(m)} + X_{(m+1)})/2 \) (we consider the case \( n = 2m \)).

    - Here we have:
      \[ T_{(n-1)}^{(i)} = \begin{cases} 
      X_{(m+1)} & \text{if } i = (1), \ldots, (m) \\
      X_{(m)} & \text{if } i = (m + 1), \ldots, (n). 
      \end{cases} \]
- So the pseudo-values are:

\[ T_{n,i} = \begin{cases} 
  nT_n - (n - 1)X_{(m+1)} & \text{if } i = (1), \ldots, (m) \\
  nT_n - (n - 1)X_{(m)} & \text{if } i = (m + 1), \ldots, (n).
\]

- The estimation of the mean of \( T_n \) is:

\[
\hat{T}_n = nT_n - \frac{m(n-1)}{n}X_{(m+1)} - \frac{m(n-1)}{n}X_{(m)}
= nT_n - (n-1)T_n = T_n
\]

- Jacknife estimation of the bias:

\[
\widehat{Bias}_{\text{jack}}(T_n) = T_n - \hat{T}_n = 0
\]

- But the jacknife estimation of the variance does not work!

- It can be shown that

\[
\widehat{Var}_{\text{jack}}(T_n) = \frac{1}{n} \hat{S}_n^2
= \frac{n-1}{4} (X_{(m+1)} - X_{(m)})^2
\]

- Whereas (for symetric densities):

\[
AVar(T_n) = \frac{1}{4n f_X^2(\theta)}
\]

- The two quantities are different even when \( n \to \infty \):

\[
\frac{1}{n} \hat{S}_n^2 - \frac{1}{4n f_X^2(\theta)} \not\to 0 \quad \text{as } n \to \infty
\]
The boundary of a support:
- Here $\theta = \inf \{ x | F(x) = 1 \}$ and $T_n = X_{(n)}$.
- One can show that
  \[ \hat{T}_n = X_{(n)} + \frac{n-1}{n} (X_{(n)} - X_{(n-1)}) \]
- One can prove that:
  \[ E(T_n) = \theta + O(n^{-1}) \]
  \[ E(\hat{T}_n) = \theta + O(n^{-2}) \]
- But here again, the jackknife estimation of the variance does not work!

7.2 The Jackknife and the Bootstrap

• Bias and variance
  - Both methods resample from the original sample $\mathcal{X}$, but in different ways.
  - We compare the formulas for estimating the bias and the variance of a statistics $T_n$ by both methods.
  - Bias:
    \[ \hat{Bias}_{jack}(T_n) = (n-1) \left[ \frac{1}{n} \sum_{i=1}^{n} T_{n-1}^{(i)} - T_n \right] \]
    \[ \hat{Bias}_{boot}(T_n) = \left[ \frac{1}{B} \sum_{b=1}^{B} T_{n}^{*(b)} - T_n \right] \]
- Variance:

\[
\widehat{\text{Var}}_{\text{jack}}(T_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (T_{n,i} - \hat{T}_n)^2
\]

\[
= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left( [nT_n - (n-1)T_{n-1}^i] - [nT_n - \frac{n-1}{n} \sum_{j=1}^{n} T_{n-1}^j] \right)^2
\]

\[
= (n-1) \left[ \frac{1}{n} \sum_{i=1}^{n} (T_{n-1}^i) - \frac{1}{n} \sum_{j=1}^{n} T_{n-1}^j \right]^2
\]

\[
\widehat{\text{Var}}_{\text{boot}}(T_n) = \frac{1}{B} \sum_{b=1}^{B} (T_{n}^{*}(b) - \frac{1}{B} \sum_{b=1}^{B} T_{n}^{*}(b))^2
\]

- Both jackknife formulas have an inflation factor: \( n - 1 \)

- this is needed because the jackknife deviations are smaller than the bootstrap deviations

- the jackknife resamples resemble much more to \( X \) than \( X^* \) do.

\begin{itemize}
  \item **Comparison of practical performances**
  \begin{itemize}
    \item Since the jackknife uses only \( n \) specific resamples from \( X \), whereas the bootstrap makes use of much more information on the data, we expect better performance of the bootstrap.
    \item But the jackknife is easier to compute (for, say, \( n \leq 200 \)).
  \end{itemize}
\end{itemize}
In fact, the jackknife can be viewed as an approximation of the bootstrap. It depends on the statistics.

- An estimator $T_n$ of $\theta$ is **linear** if
  
  $$T_n = \frac{1}{n} \sum_{i=1}^{n} \alpha(X_i),$$

  for some function $\alpha$ (example: sample mean).

- An estimator $T_n$ of $\theta$ is **quadratic** if
  
  $$T_n = \frac{1}{n} \sum_{i=1}^{n} \alpha(X_i) + \frac{1}{n^2} \sum_{i<j} \beta(X_i, X_j),$$

  for some functions $\alpha$ and $\beta$ (example: sample variance).

**Estimation of the variance**

- If $T_n$ is linear:
  
  - One can show:
    
    $$\hat{\text{Var}}_{\text{jack}}(T_n) = \frac{n}{n-1} \hat{\text{Var}}_{\text{boot}}(T_n)$$

    where $\hat{\text{Var}}_{\text{boot}}(T_n)$ is the true bootstrap variance (computed with the ideal $B = \infty$)

  - Example: sample mean, $T_n = \bar{X}_n$, we have
    
    $$\hat{\text{Var}}_{\text{jack}}(T_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
    
    $$\hat{\text{Var}}_{\text{boot}}(T_n) = \frac{1}{n^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
- If $T_n$ is not linear:
  - Let $T_{\text{lin},n} = (1/n) \sum_{i=1}^{n} \alpha(X_i)$ be a linear approximation of $T_n$, that equals $T_n$ for the $n$ jackknife resamples $X^{(i)}$ (system of $n$ equations in $n$ unknowns $\alpha_i = \alpha(X_i)$).
  - Then
    $$\hat{\text{Var}}_{\text{jack}}(T_n) = \frac{n}{n-1} \hat{\text{Var}}_{\text{boot}}(T_{\text{lin},n})$$
  - The quality of the jackknife estimator depends on the quality of the linear approximation (if $T_n$ highly nonlinear, the jackknife could be inefficient).

**Estimation of the bias**

- If $T_n$ is quadratic:
  - It can be shown that
    $$\hat{\text{Bias}}_{\text{jack}}(T_n) = \frac{n}{n-1} \hat{\text{Bias}}_{\text{boot}}(T_n)$$
  - Example: plug-in variance $T_n = (1/n) \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$
    $$\hat{\text{Bias}}_{\text{jack}}(T_n) = T_n - \hat{T}_n$$
    $$= \left(\frac{1}{n} - \frac{1}{n-1}\right) \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
    $$= -\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
    $$\hat{\text{Bias}}_{\text{boot}}(T_n) = -\frac{1}{n^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$
- If $T_n$ is not quadratic:

$$\hat{\text{Bias}}_{\text{jack}}(T_n) = \frac{n}{n - 1} \hat{\text{Bias}}_{\text{boot}}(T_{\text{quad},n}),$$

where $T_{\text{quad},n}$ is a quadratic approximation of $T_n$ that equals $T_n$ for the jackknife samples.

- **Example:** Monte-Carlo comparison of jackknife and bootstrap for estimating the standard deviation of a statistics.

  - We simulate 1000 samples of size $n = 10$ from a bivariate normal distribution with $\mu_1 = \mu_2 = 1$, $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.7$.

  - We analyze and compare the jackknife and the bootstrap estimates ($B = 1000$) of **standard error** of 3 statistics:

    - $T_1 = \bar{X}_1$
    - $T_2 = r_{12} = \frac{S_{12}}{S_1 S_2}$
    - $T_3 = (\bar{X}_1)^2$

    - $T_1$ is linear, but $T_2$ and $T_3$ are not.

    - Figure 7.1 displays the boxplots of the estimates for the 3 statistics, over the 1000 simulations.

    - The results of both methods are quite similar but with a slightly greater dispersion for the jackknife estimates for the nonlinear statistics $T_2$ and $T_3$.

    - Note that the true value for $\text{Std}(T_1) = \frac{\sigma}{\sqrt{n}} = 0.3162$. 


7.3 The Jackknife and the Delta method

- The jackknife can be used to approximate the empirical influence function evaluated at the observed $x_i$.

  - Let $T_n = T(X) = \theta(F_n)$ be a plug-in estimator of $\theta(F)$. The influence function of $\theta$ is:
    \[
    L_\theta(x; F) = \frac{\partial \theta((1 - \varepsilon)F + \varepsilon H_x)}{\partial \varepsilon} \bigg|_{\varepsilon=0}.
    \]

  - The nonparametric delta method provides:
    \[
    \theta(F_n) - \theta(F) \sim AN(0, \text{var}_L(F)).
    \]
– In practice, the variance is approximated by plugging $F_n$ in place of $F$:

$$\text{var}_L(F_n) = \frac{1}{n^2} \sum_{i=1}^{n} L^2_\theta(x_i; F_n)$$

– A numerical approximation to the derivative $L_\theta(x_i; F_n)$ is obtained through

$$L_\theta(x_i; F_n) = \frac{\theta ((1 - \varepsilon)F_n + \varepsilon H x_i) - \theta(F_n)}{\varepsilon}$$

where $\varepsilon \to 0$ when $n \to \infty$.

– It can be shown that when choosing $\varepsilon = -(n - 1)^{-1}$, we obtain the jackknife approximation of $L_\theta(x_i; F_n)$:

$$L_\theta(x_i; F_n) \approx (n - 1) \left( T_n - T^{(i)}_{n-1} \right),$$

so we have

$$\text{var}_L(F_n) \approx \frac{n - 1}{n^2} \sum_{i=1}^{n} \left( T_n - T^{(i)}_{n-1} \right)^2.$$
Chapter 8

The Smoothed Bootstrap

- For the nonparametric bootstrap, we used mostly $F_n$ as an estimator of $F$. But $F_n$ is discrete.

  - The discreteness of $F_n$ could create additional noise when the true $F$ is continuous.

  - Smoothing would be suitable for statistics like a median, a quantile or a boundary of $X$, where discreteness draws too many undesirable values in the bootstrap sample (like $X_{(n)}$) when estimating the upper boundary of the support of $X$.

  - It can be shown, for example, that for the median, the smoothed bootstrap, with $h \propto n^{-1/5}$, can achieve $O(n^{-4/5})$ accuracy, in place of the usual $O(n^{-1/2})$ for the ordinary bootstrap.

8.1 Smooth Estimates of $F$

- A standard smooth estimate of a continuous pdf is the **kernel density estimation** (see Silverman, 1986).
we start with a series of \( n \) univariate observed values of \( x \): \( x_1, \ldots, x_n \)

- the discrete density estimator put a mass \( 1/n \) at each observed data point. We can write the discrete density as

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x - x_i),
\]

where \( I(\cdot) \) is a dirac-type function putting a mass 1 at zero, \( I(z) = 1 \), if and only if \( z = 0 \) and \( I(z) = 0 \) elsewhere.

- the idea of Kernel smoothing is to replace this discontinuous dirac-type function by a continuous density \( K(\cdot) \) centered at zero, as \( I(\cdot) \), with a dispersion controled by a tuning parameter \( h \), called the bandwidth.

- The kernel density estimator is thus given by:

\[
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),
\]

- \( K(\cdot) \) is a given function called the Kernel and \( h \) is the a given smoothing parameter called the bandwidth.

  - Gaussian Kernel: \( K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2 \right) \), for \( u \in \mathbb{R} \);
  - Quartic Kernel: \( K(u) = \frac{15}{16}(1 - u^2)^2 \), for \( -1 \leq u \leq 1 \)

- \( \hat{f}_h(x) \) is the average of \( n \) densities (standard gaussian if we use a gaussian kernel) centered at the observed points \( x_i \). The bandwidth \( h \) is a tuning parameter which control the dispersion of these gaussian densities.
- if $h$ is small only “local” observations (near the point $x$) will influence $\hat{f}_h(x)$
- if $h$ is very large, all the observations will play a role in the computation of $\hat{f}_h(x)$.
- if $h \to 0$, the density will degenerate into the empirical density which is the discrete density giving a mass $1/n$ at each observed point (no smoothing at all)
- if $h \to \infty$, the density will degenerate in a flat horizontal line (too smoothed).

- **Example:** A sample of size $n = 20$ was simulated from a normal variate: $X \sim N(10, 2)$. Figure 8.1 displays the kernel density estimate (the bandwidth $h = 1.0822$, rule of thumb) and the true normal density.

![Figure 8.1: Kernel density estimator: the true density is the dash-dotted line, the kernel estimate is the solid line. The small gaussian densities on the bottom are the kernel functions centered on the $n$ data points and scaled by $n$: they are added to produce the estimator.](image)
• Choice of the bandwidth $h$

- The quality of the density estimator is not very sensitive to the choice of the kernel function $K$. However, as pointed above the choice of the smoothing parameter, the bandwidth $h$, is crucial.

- Theoretical results exist and practical rules based on the data are available:
  
  - Theoretical result: under regularity conditions on the true $f$ and on $K$, and for a good choice of $h = O(n^{-1/5})$, $\hat{f}_h(x) \to f(x)$ as $n \to \infty$.

  - Normal reference rule (Rule of thumb): if the true density is not too far from normality, the optimal value for $h$ is:
    
    For Gaussian Kernel: $h_G = 1.06 \hat{\sigma} n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation.

    For Quartic Kernel: $h_Q = 2.62 h_G$.

  - Robust rule of thumb: a more robust optimal value for $h$ is:
    
    For Gaussian Kernel: $h_R = 1.06 \min \left( \hat{\sigma}, \frac{\hat{R}}{1.349} \right) n^{-1/5}$, where $\hat{R}$ is the interquartile range ($\hat{R} = q_{0.75} - q_{0.25}$).

  - Other data driven method for choosing $h$ are available. An easy rule is based on cross-validation (see Silverman, 1986): minimization of an estimator of the integrated mean squared error (estimation obtain by a leave-on-out technique).
8.2 The bootstrap algorithm

- The idea: resampling from $\hat{F}_h$ in place of $F_n$ where $\hat{F}_h$ is the continuous cdf corresponding to the pdf $\hat{f}_h$.

- Consider the random sample $\mathcal{X}_s^* = \{X_{1,s}^*, \ldots, X_{n,s}^*\}$

$$X_{i,s}^* = X_i^* + h \varepsilon_i,$$

where $X_i^* \sim F_n$ is an usual random drawn from $\mathcal{X}$, and $\varepsilon_i \sim K(\cdot)$ is a random drawn from the kernel density $K$.

- It is easy to prove that $\mathcal{X}_s^*$ is a random sample of size $n$ from $\hat{F}_h$: $X_{i,s}^* \sim \hat{F}_h$.

- This is particularly easy to perform if we use a Gaussian kernel.

- Variance correction:

- We see that $Var(X_s^*) = Var(X^*) + h^2 \neq Var(X^*)$, the plug-in estimate of the variance of the data.

- We introduce the following correction:

$$X_{i,s,corr}^* = \bar{X}^* + (1 + h^2/Var(X^*))^{-1/2} (X_i^* - \bar{X}^* + h \varepsilon_i)$$

- $X_{i,s,corr}^*$ has an expectation $E(X^*)$ and a variance approximately equal to $Var(X^*)$

- For each bootstrap sample $\mathcal{X}_{s,corr}^*$, we compute the value of the statistics of interest $T(\mathcal{X}_{s,corr}^*)$ and we can use the bootstrap algorithm as usual.
• Example:

- Estimating the 0.95 quantile of an exponential from a random sample $X_i \sim \text{Expo}(\mu = 10)$ with $n = 100$.
  
  - The true value $q_{0.95} = 29.9573$ and the estimate is $\hat{q}_{0.95} = 36.6057$.
  
  - Chosen bandwidth for estimating the density of $X$ is $h = 3.8761$ (robust rule of thumb).

Figure 8.2: Monte-Carlo distribution of $\hat{q}_{0.95}$ with naive bootstrap. Here $B = 5000$.

Figure 8.3: Monte-Carlo distribution of $\hat{q}_{0.95}$ with smoothed bootstrap with $h = 3.8761$. Here $B = 5000$. 
8.3 Testing Unimodality

- The smoothed bootstrap can be used to test the unimodality of a density.
  - The idea: if $h$ decreases, the number of modes of $\hat{f}_h(x)$ increases.
  - We test unimodality by seeing if unusually large values of $h$ is needed to make $\hat{f}_h(x)$ unimodal.
    - Test statistics:
      $$ T = \min_{h > 0} \{ h \mid \hat{f}_h(x) \text{ is unimodal} \} $$
    - We reject if $T$ is too large. Let $t = T_{observed}$.
    - The $p$-value is
      $$ p - \text{value} = \text{Prob}(T \geq t \mid H_0) $$
      where $H_0$ is the unimodality of $f(x)$.

- The algorithm:
  -[1 ] we generate $B$ samples $X_{s,corr}^{*,b}$, $b = 1, \ldots, B$ under the null, i.e. drawn from $\hat{f}_t(x)$.
  -[2 ] For each sample, we compute $T^{*,b}$
  -[3 ] The estimated $p$-value is:
    $$ \hat{p} = \frac{\#\{T^{*,b} \geq t\}}{B} $$
    - The steps [2]-[3] are more easily computed by computing $\hat{f}_t^*(x)$ for the bootstrap sample: the event $T^{*,b} \geq t$ happens if and only if $\hat{f}_t^*(x)$ shows more than one mode.
Chapter 9

The Bootstrap for Time Series: an introduction

- The observations are time dependent: we cannot resample from the empirical cdf $F_n$ which does not reflect the time dependency.
  - Two main approaches:
    - Using a parametric model for resampling
    - Using the blocking of the data to simulate the time series.

9.1 Model based resampling

- We use standard time series models (ARMA) for defining the data generating process.
  - Let $\{X_t\}$ be a second order stationary time series, with zero mean and autocovariance function $\gamma_k$: for all $t, k$ we have
    $$E(X_t) = 0$$
    $$Cov(X_t, X_{t+k}) = \gamma_k.$$ 
  - The autocorrelation function is $\rho_k = \gamma_k/\gamma_0$ for all $k = 0, \pm 1, \pm 2, \ldots$
Some basic models

- MA(1):
  * The model
    \[ X_t = \varepsilon_t + \beta \varepsilon_{t-1}, \quad t = \ldots, -1, 0, 1, \ldots \]
  * here \( \{\varepsilon_t\} \) is a white noise process of innovations:
    \[ \varepsilon_t \sim (0, \sigma^2), \text{ independent} \]
  * the autocorrelation function is \( \rho_1 = \beta/(1+\beta^2) \) and \( \rho_k = 0 \) for \( |k| > 1 \).

- AR(1):
  * The model
    \[ X_t = \alpha X_{t-1} + \varepsilon_t, \quad t = \ldots, -1, 0, 1, \ldots \]
  * here \( |\alpha| < 1 \), and \( \{\varepsilon_t\} \) is a white noise
  * the autocorrelation function is \( \rho_k = \alpha^{|k|} \).

- ARMA(p,q):
  * The model
    \[ X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t + \sum_{k=1}^{q} \beta_k \varepsilon_{t-k}, \]
    \[ \{\varepsilon_t\} \text{ is a white noise.} \]
  * Conditions on the coefficients to obtain a stationary process.
• The bootstrap

- Same idea as in the regression models
  - Fit the model to the data
  - Constructed the residuals from the fitted model
  - Recenter the residuals (mean zero as $\varepsilon_t$)
  - Generate new series by incorporating random samples from the fitted residuals in the fitted model

- Example: AR(1) model.
  - We have a sample $x_1, \ldots, x_n$, we compute $\hat{\alpha}$: we must have $|\hat{\alpha}| < 1$.
  - Estimated innovations (residuals):
    \[ e_t = x_t - \hat{\alpha} x_{t-1}, \quad \text{for } t = 2, \ldots, n \]
    - $e_1$ is unavailable because $x_0$ unknown.
    - recenter the residuals $\tilde{e}_t = e_t - \bar{e}$
    - draw with replacement from the set $\tilde{e}_t, t = 2, \ldots, n$ to obtain the $n + 1$ bootstrap innovations $\varepsilon_0^*, \ldots, \varepsilon_n^*$
    - Then we define the bootstrap sample $X_0^*, \ldots, X_n^*$:
      \[ X_0^* = \varepsilon_0^* \]
      \[ X_t^* = \hat{\alpha} X_{t-1}^* + \varepsilon_t^*, \quad t = 1, \ldots, n \]
    - Then compute $\hat{\alpha}^*$ from the observation $X_1^*, \ldots, X_n^*$, and proceed as usual.
- In fact the series \( \{X_t^*\} \) is not stationary. For improving this:

- Start the series of \( X_t^* \) at \( t = -k \) (in place of \( t = 0 \)) and then discard the observations \( X_{-k}^*, \ldots, X_0^* \).
- If \( k \) is large enough, the resulting bootstrap sample will be approximatively stationary.

- Can be used for testing hypothesis, confidence intervals (bootstrap-t, percentile,\ldots), prediction (see Thombs and Schucany, 1990),\ldots

## 9.2 Block resampling

* The idea here is not to draw from innovations defined with respect to a particular model but on blocks of consecutive observations.

- We don’t need a model here...The algorithm is as follows:

  - Divide the data into \( b \) non-overlapping blocks of length \( \ell \). So we suppose \( n = b\ell \).
  - Define the blocks \( y_j \):
    \[
    y_1 = (x_1, \ldots, x_\ell) \\
    y_2 = (x_{\ell+1}, \ldots, x_{2\ell}) \\
    \vdots \\
    y_b = (x_{(b-1)\ell+1}, \ldots, x_{b\ell})
    \]
  - Take a bootstrap sample of \( b \) blocks drawn independently with replacement from the set \( (y_1, \ldots, y_b) \).
- This produce a bootstrap sample \((y_1^*, \ldots, y_b^*)\) which gives a bootstrap series of length \(n\).

  - The idea is that if \(\ell\) is large enough, the original dependence in the series is preserved

- So we hope that the bootstrap statistics \(T(\mathcal{X}^*)\) will have approximatively the same distribution as the statistics \(T(\mathcal{X})\) in the real world, with the real series...

- On the other hand we must have enough blocks \(b\) to estimate with accuracy the bootstrap distribution of \(T(\mathcal{X}^*)\)

  - A good balance: \(\ell\) should be of the order \(O(n^\gamma)\) where \(\gamma \in (0, 1)\), so \(b = n/\ell \to \infty\) when \(n \to \infty\).

  - There are several variants of the method (overlapping blocks, blocks of blocks, blocks of random length, pre-withening, ...).

- Theoretical works still in progress (optimal choice of \(\ell\), etc...):

- See Lahiri (1999), Politis and Romano (1994), Politis and White (2003), Politis, Romano and Wolf (2001),...
Chapter 10

Other Simulation Based Estimation Methods

• There exist other estimations techniques using simulation which are not related to the bootstrap ideas presented here.
  – Maximum Simulated Likelihood (MSL)
  – Method of Simulated Moments (MSM)
  – Method of Simulated Scores (MSS)

• All these methods are adapted to problems (essentially parametric models) where MLE (Maximum Likelihood Estimation) and GMM (Generalized Method of Moments) are useful.
  – The likelihood function and/or the moment (or score) function, are often difficult to evaluate numerically
    - high dimensional integrals
    - mixed discrete and continuous densities
    - evaluation of multidimensional probabilities
Often these quantities can be expressed as expectation of random variables which, for a given value of the parameters, can be easily simulated.

The idea: use Monte-Carlo simulations to evaluate these expectations, by taking the appropriate empirical mean of the simulated values.

For example: \( \int_0^1 f(x)\,dx = E(f(X)) \) where \( X \sim U(0,1) \), so

\[
\int_0^1 f(x)\,dx \approx \frac{1}{M} \sum_{i=1}^M f(X_i)
\]

where \( X_i \) are random drawn from \( U(0,1) \).

Maximization and/or moment (or score) condition are then numerically handled by using these Monte-Carlo approximations.

- All these methods are very specific to the particular problem analyzed.
  - Censoring and truncating
  - Probit and Tobit models
  - More generally: limited dependent variables models.
• **Example** Probit models.

- The standard binomial probit model is easy:
  
  - we observe a dichotomous variable $Y$ which is a *censored* random variable of a latent unobserved $Y^*$ variable, explained by regressors $X$.

  $\begin{align*}
  Y_i &= \begin{cases} 
  0 & \text{if } Y^*_i < 0 \\
  1 & \text{if } Y^*_i \geq 0 
  \end{cases} 
  \text{ where } Y^*_i | X_i = x_i \sim N(x_i' \beta, 1) 
  \end{align*}$

  - Equivalently

    \[
    P(Y_i = 1 | X_i = x_i) = \Phi(x_i' \beta, 1) \\
    f(y_i | x_i) = (\Phi(x_i' \beta, 1))^{y_i} (1 - \Phi(x_i' \beta, 1))^{1-y_i} I_{\{0,1\}}(y_i)
    \]

  - The Likelihood is easy to compute if independence:

    \[
    L(\beta; y_1, \ldots, y_n) = \prod_{i=1}^{n} \left[ (\Phi(x_i' \beta, 1))^{y_i} (1 - \Phi(x_i' \beta, 1))^{1-y_i} \right]
    \]

  - The multinomial probit is more complicated: we have $p$ possible outcomes.

  - The latent variable $Y^*$ is $N_p(\eta, \Omega)$ where the elements of $\eta$ are $\eta_k = x' \beta_k$, $k = 1, \ldots, p$ for some explanatory variables $x$ and some vector $\beta_k$ of appropriate dimension.

  - The observed $Y$ is the censored variable defined as the indicator for the maximal element of $Y^*$:

    \[
    Y = I_{\{\max_i Y_i^*\}}(Y^*)
    \]
- The sample space $S$, for $Y$ is the set of rows $S_j$ of the identity matrix $I_p$: $S_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the $j$th column.

- The probability function for $Y$ is now written as a $p - 1$ dimensional integral:

$$P(Y = S_j) = P(Y_j = 1, Y_k = 0, k \neq j)$$

$$= P(Y_j^* - Y_k^* \geq 0, k = 1, \ldots, p, k \neq j)$$

* defining $Z_j = \{Y_j^* - Y_k^*, k = 1, \ldots, p, k \neq j\} = D_j Y^*$,

where $D_j$ is the appropriate $(p - 1) \times p$ first difference operator.

* So, $Z_j \sim N_{p-1}(\mu_j, \Omega_j)$ where $\mu_j = D_j \eta$ and $\Omega_j = D_j \Omega D_j'$.

- we obtain

$$f(y; \theta) = \begin{cases} 
\prod_{j=1}^{p} \Phi_{p-1}(-\mu_j, \Omega_j)^{y_j} & \text{for } y \in S \\
0 & \text{otherwise.}
\end{cases}$$

where for notation, the general expression $\Phi_p(\mu, \Omega)$ is the $p$-dimensional integral:

$$\Phi_p(\mu, \Omega) = \int_{-\infty}^{0} \cdots \int_{-\infty}^{0} \phi_p(x + \mu, \Omega) \, dx$$

- This is numerically untractable when $p$ large (say, $p \geq 4$).

- But this is easy to simulate:

- For any value of $\theta = (\eta, \Omega)$ we can simulate $Y^*^{(b)} \sim N_p(\eta, \Omega)$ a large number of time, $b = 1, \ldots, B$. 


- We approximate $f(y; \theta)$ for any value $y \in S$, by the empirical Monte-Carlo proportions

- For instance, for $y_i = s_j$ we have:

$$\hat{P}(Y = s_j) = \hat{f}(y_i; \eta, \Omega)$$

$$= \hat{P}(Y_j^{*^{(b)}} - Y_k^{*^{(b)}} \geq 0, k = 1, \ldots, p, k \neq j)$$

$$= \frac{\#\{Y_j^{*^{(b)}} - Y_k^{*^{(b)}} \geq 0, k = 1, \ldots, p, k \neq j\}}{B}.$$ 

- Then we can compute the likelihood function for any values of $\eta, \Omega$

$$\ln L(\eta, \Omega; y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln \hat{f}(y_i; \eta, \Omega)$$

- Note that $P(Y_j^{*} - Y_k^{*} \geq 0, k = 1, \ldots, p)$ can be written as an expectation of an appropriate indicator function in terms of $Y^*$.

- Similar expressions (expectations of functions of the variables $Y^*$) can be determined for the score function itself or for any other general moment functions.

- The approach is quite different than the bootstrap ideas...and the computations are really dependent on the specific particular problem analyzed.

• **In the bootstrap**, we do not look for defining an estimator, but rather to analyze the sampling properties of a given estimator.

  - The Monte-Carlo technique is used to approximate these sampling distribution.
    - to compute bias and variance of the estimator
    - to build confidence intervals
    - to compute $p$-values, . . .

• Recently some works propose to use the bootstrap as a tool for improving the quality of estimators in nonlinear problems (like most GMM estimators)

  - to improve the statistical accuracy of an estimator (“bagging” = bootstrapping and averaging), (Breiman, 1996)
  
  - to stabilize numerically some estimator obtained through highly nonlinear optimization procedure (Climov, Delecroix and Simar, 2001)
Chapter 11

Is the Bootstrap a Panacea?

11.1 Some Conclusions

• Good points
  – Bootstrap often works and it is easy to implement
  – Might offer better approximations than usual first order asymptotics
  – Often, it is the only way to construct CI, to correct for the bias of an estimator or to estimate standard errors of estimates.
  – It is very flexible

• Bad points
  – May be computer intensive
  – May not work (not being consistent)

• Many other applications
  – Nonlinear regression models
– The Bootstrap in Nonparametric regression models
– Nonparametric statistics
– Advanced time series
– Subsampling
– Bootstrap for extremes
– Censored data
– ...(see the literature)

11.2 Personal and Quick Insights in the Literature

• Textbooks with broad spectrum

• Introductory surveys

• Historical references
Quenouille (1949), Efron (1982)

• Theoretical references

• Various Applications


– Panel data: Hall, Härdle and Simar (1995)


– Smoothed bootstrap: Silverman and Young (1987)


– See the list below, . . .
Bibliography


Appendix A: Gamma and related distributions

- Some reminders and notations
  - Gamma distribution: $X \sim \Gamma(\alpha, \beta)$
    \[
    E(X) = \alpha \beta \\
    V(X) = \alpha \beta^2 \\
    f(x) = \frac{x^{\alpha-1} \exp(-x/\beta)}{\beta^\alpha \Gamma(\alpha)}
    \]
  - Exponential distribution: $X \sim \text{Expo}(\theta) \equiv \Gamma(1, \theta)$
    \[
    E(X) = \theta \\
    V(X) = \theta^2 \\
    f(x) = \frac{\exp(-x/\theta)}{\theta}
    \]
  - Chi-square distribution: $X \sim \chi^2_\nu \equiv \Gamma(\nu/2, 2)$.
    \[
    E(X) = \nu \\
    V(X) = 2\nu
    \]
• Normal sampling

- \( X_i \sim \mathcal{N}(\mu, \sigma^2), \ i = 1, \ldots, n \). We have:

\[
\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2_{n-1}
\]

\[
S^2 \sim \frac{\sigma^2}{n - 1} \chi^2_{n-1} \equiv \Gamma\left(\frac{n - 1}{2}, \frac{2\sigma^2}{n - 1}\right)
\]

• Exponential sampling

- \( X_i \sim \text{Exp}(\mu), \ i = 1, \ldots, n \). We have:

\[
\bar{X} \sim \Gamma(n, \frac{\mu}{n})
\]

\[
\frac{\bar{X}}{\mu} \sim \Gamma(n, \frac{1}{n}).
\]