INSTITUT DE STATISTIQUE
Université Catholique de Louvain
Louvain-la-Neuve
BELGIUM

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

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): $\mathcal{X} = (X_1, \ldots, X_n)$ 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 $\tau \in I\!R^k$ 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 $T(\mathcal{X})$ is an estimator of $\theta(F)$. Examples:

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

• **Measures of the accuracy of the estimator $T(\mathcal{X})$?**

Quantities of interest:

$$Bias_F(T) = E_F T(\mathcal{X}) - \theta(F)$$
$$Var_F(T) = E_F T^2(\mathcal{X}) - (E_F T(\mathcal{X}))^2$$
$$MSE_F(T) = E_F((T(\mathcal{X}) - \theta(F))^2)$$
$$G(x) = Dist_{T,F}(x) = P_F(T(\mathcal{X}) \leq x)$$

• 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:

$$\begin{aligned} H(x) &= P_F(T(\mathcal{X}) - \theta(F) \leq x) \\ \text{where} \quad H(x) &= Dist_{T,F}(x + \theta(F)) = G(x + \theta(F)) \end{aligned}$$

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

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

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

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

– 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:

$$\begin{aligned} R1_n &= \sqrt{n}(\bar{X} - \mu) \sim N(0, \sigma^2) \\ R2_n &= \sqrt{n}\frac{(\bar{X} - \mu)}{S} \sim t_{n-1} \end{aligned}$$

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

– Confidence interval for μ using $R1_n$ **if** σ **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 μ using $R2_n$ **if** σ **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 σ is unknown, $R1_n$ is a root and $R2_n$ is a pivotal root.

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

– Only asymptotic results

– $R1_n$ is a root with $R1_n \sim AN(0, \sigma^2)$

– $R2_n$ is an asymptotically pivotal root : $R2_n \sim AN(0, 1)$.

– **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

$$G(x) = Dist_{T,F}(x) = P_F(T(\mathcal{X}) \leq x).$$

  - **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 :

$$
\begin{aligned}
Bias^*(T) &= Bias_{F_n}(T) = E_{F_n} T(\mathcal{X}^*) - \theta(F_n) \\
Var^*(T) &= Var_{F_n}(T) = E_{F_n} T^2(\mathcal{X}^*) - \left(E_{F_n} T(\mathcal{X}^*)\right)^2 \\
MSE^*(T) &= MSE_{F_n}(T) = E_{F_n}\left([T(\mathcal{X}^*) - \theta(F_n)]^2\right) \\
Dist_T^*(x) &= Dist_{T,F_n}(x) = P_{F_n}\left(T(\mathcal{X}^*) \leq x\right) = P\left(T(\mathcal{X}^*) \leq x \,|\, \mathcal{X}\right)
\end{aligned}
$$

• In particular, $\widehat{G}(x) = Dist_{T,F_n}(x)$ is the bootstrap estimate of $G(x) = Dist_{T,F}(x) = P_F(T(\mathcal{X}) \leq x)$

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

$$
\begin{aligned}
\widehat{H}(x) &= P_{F_n}\left(T(\mathcal{X}^*) - \theta(F_n) \leq x\right) \\
&= P\left(T(\mathcal{X}^*) - \theta(F_n) \leq x \,|\, \mathcal{X}\right) \\
&= \widehat{G}(x + \theta(F_n))
\end{aligned}
$$

• **The Key relation**:
under "regularity conditions" (see below), when n is large:

$$\widehat{G}(x) = Dist_{T,F_n}(x) \approx Dist_{T,F}(x) = G(x)$$

• **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 $\tau \in I\!R^k$ 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) τ

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

$$\Rightarrow \underline{\text{Monte-Carlo approximations}}$$

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

• **Example**: Nonparametric Bootstrap, Estimation of a mean μ 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**: $X \sim F(\cdot)$

$$E_F(X) = \mu$$
$$Var_F(X) = \sigma^2$$

– We have in the **bootstrap world**: $X^*|\mathcal{X} \sim F_n(\cdot)$

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

$$\begin{aligned} Var^*(X^*) &= Var_{F_n}(X^*) = E_{F_n}(X^* - E_{F_n}(X^*))^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n-1}{n} S^2 \end{aligned}$$

– 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, $Bias_F(T)$, $Var_F(T)$, $MSE_F(T)$ and $Dist_{T,F}(x)$ could eventually be calculated by analytical methods.
  - Example: If $X_i \sim$ Expo (μ) with $\mu = E(X_i)$ (see the Appendix A). We chose the statistic $T(\mathcal{X}) = \bar{X}$:
    - then $\bar{X} \sim \Gamma(n, \mu/n)$

$$G(x) = P_F(\bar{X} \leq x) = Dist_{T,F}(x) = \int_0^x \frac{u^{n-1}\exp(-nu/\mu)}{(\mu/n)^n\Gamma(n)}\, du$$

    - if n is large, then $\bar{X} \sim AN(\mu, \mu^2/n)$

$$G(x) = P_F(\bar{X} \leq x) = Dist_{T,F}(x) \approx \Phi\left(\frac{x-\mu}{\mu/\sqrt{n}}\right)$$

  - $G(x)$ 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

$$G(x) = Dist_{T,F}(x) \approx \frac{1}{B}\sum_{b=1}^{B} I(T(\mathcal{X}^{(b)}) \leq x)$$

– 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:

$$Bias_F(T) \approx \frac{1}{B}\sum_{b=1}^{B} T(\mathcal{X}^{(b)}) - \theta(F)$$

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

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

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

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

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

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

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

- 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.

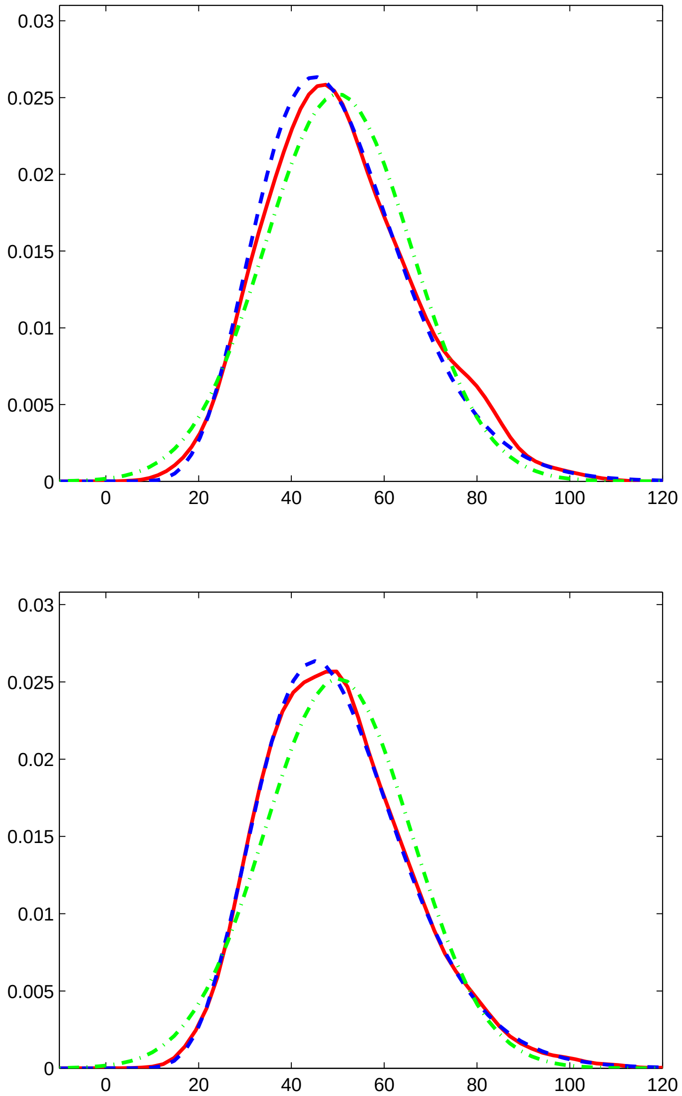

Figure 1.1: *Sampling distribution of* $\bar{X}$, $X_i \sim$ *Expo* $(\mu = 50)$ *with* $n = 10$. *True Gamma (dashed), Normal approximation (dashdot) and MC approximation (solid): top* $B = 500$ *and bottom* $B = 5000$.

• 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:

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

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

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

$$\widehat{G}(x) = Dist_T^*(x) \approx \frac{1}{B}\sum_{b=1}^{B} I(T(\mathcal{X}^{*(b)}) \leq x)$$

- 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 $X^* \sim Expo(\bar{x})$.

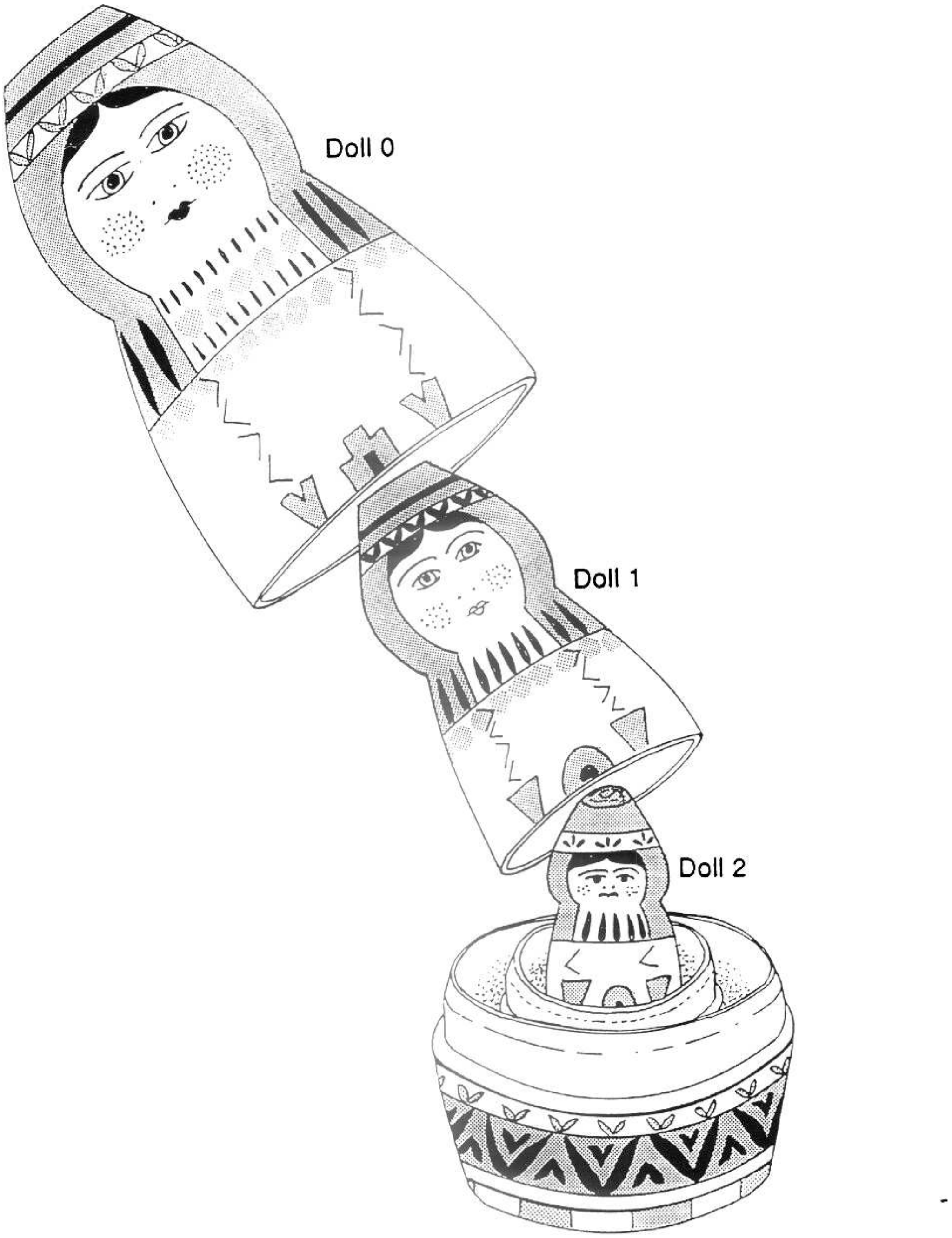

Figure 1.2: *The bootstrap analogy: the russian dolls (from Hall, 1992).*

• **Example:** Nonparametric bootstrap $X \sim (\mu, \sigma^2)$, $n = 5$
The DGP is completely unknown: we observe the data,

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

```
    original data vector            xbar        S^2
    22    10    15    18    20      17.00       22.00

    bootstrap data vectors          xbar*  for b=1,10
    10    10    18    18    20      15.20
    10    20    10    18    20      15.60
    20    22    10    22    15      17.80
    15    22    22    15    15      17.80
    20    10    15    22    15      16.40
    20    20    20    18    10      17.60
    15    18    15    18    15      16.20
    15    22    18    20    15      18.00
    20    20    18    15    18      18.20
    22    20    20    18    10      18.00
     .     .     .     .     .       .
     .     .     .     .     .       .

bootstrap mean and bootstrap variance of xbar* for B=1000
        17.0504             3.3949
```

Sampling distribution: see figure 1.3.

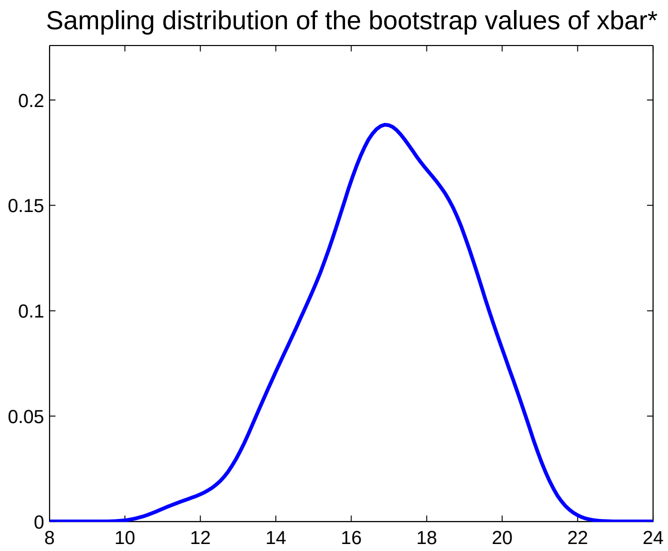

Figure 1.3: $\widehat{G}(x)$, *the bootstrap sampling distribution of* $\bar{X}$,$n = 5$ *and* $B = 1000$.

• Example: Bootstrapping the correlation

Let a sample be generated from a bivariate Normal distribution

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

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

- We want the **sampling distribution of** $\hat{\rho}(\mathcal{X})$
- Let a sample of size $n = 100$ giving $\hat{\mu} = (-0.1077 \;\; -0.0908)'$ and $S = \begin{pmatrix} 4.2567 & 1.3662 \\ 1.3662 & 0.9974 \end{pmatrix}$. So, $r = 0.6631$.
- **no real alternatives** here to estimate the sampling distribution of $\hat{\rho}(\mathcal{X})$
  (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 $Var(\hat{\rho})$:

$$\begin{aligned} E^*(\hat{\rho}) &= (1/5000) \sum_{b=1}^{5000} \hat{\rho}(\mathcal{X}^{*,b}) \\ &= 0.6622 \\ Var^*(\hat{\rho}) &= (1/5000) \sum_{b=1}^{5000} \hat{\rho}^2(\mathcal{X}^{*,b}) - \left((1/5000) \sum_{b=1}^{5000} \hat{\rho}(\mathcal{X}^{*,b}) \right)^2 \\ &= 0.0023. \end{aligned}$$

– Figure 1.4 displays the Bootstrap distribution of $\hat{\rho}(\mathcal{X}^*)$

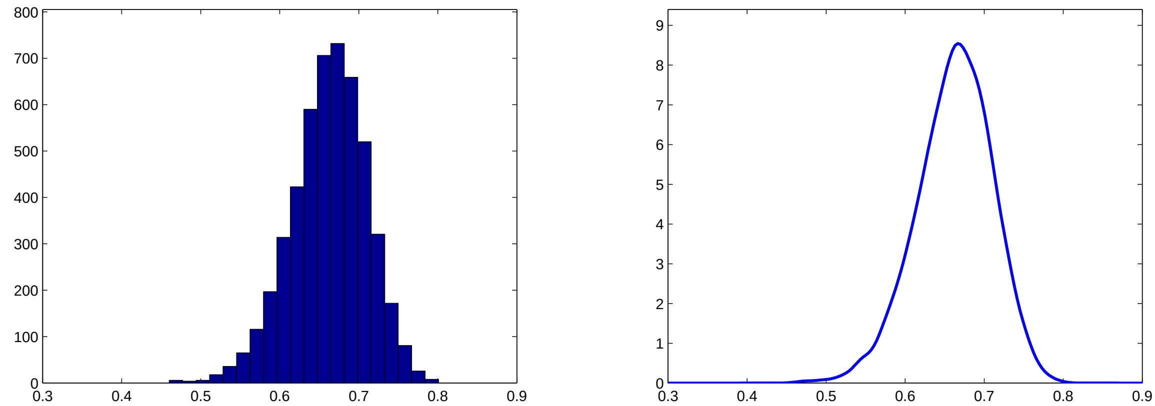

Figure 1.4: *Nonparametric Bootstrap approximation,* $B = 5000$ *of the sampling distribution of* $\hat{\rho}(\mathcal{X})$. *Left panel: histogram, right panel smoothed density estimate.*

• **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: $\mu = \sigma$).

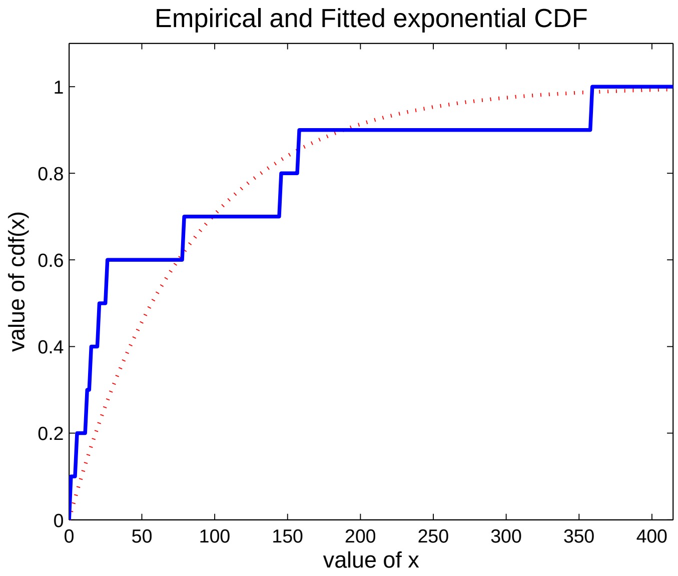

Figure 1.5: *Empirical cdf and exponential cdf of Expo*($\bar{x}$).

We are interested in the sampling distribution of $T(\mathcal{X}) = \bar{X}$, as potential estimator of μ. 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
$$\bar{X} \sim \Gamma(n, \frac{\bar{x}}{n})$$
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 $\mathcal{X}^{*(b)}$ of size $n = 10$ drawn from $Expo(\bar{x})$ and compute for each $\bar{X}^{*(b)}$. The 5000 values $\bar{X}^{*(b)}, b = 1, 5000$ provide the emprical 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 $\mathcal{X}^{*(b)}$ of size $n = 10$ drawn from $\mathcal{X}$ and compute for each $\bar{X}^{*(b)}$. The 5000 values $\bar{X}^{*(b)}, b = 1, 5000$ provide the emprical bootstrap approximation of the sampling distribution of $\bar{X}$. This result is always available.

Figure 1.6 shows the 5 different corresponding densities.

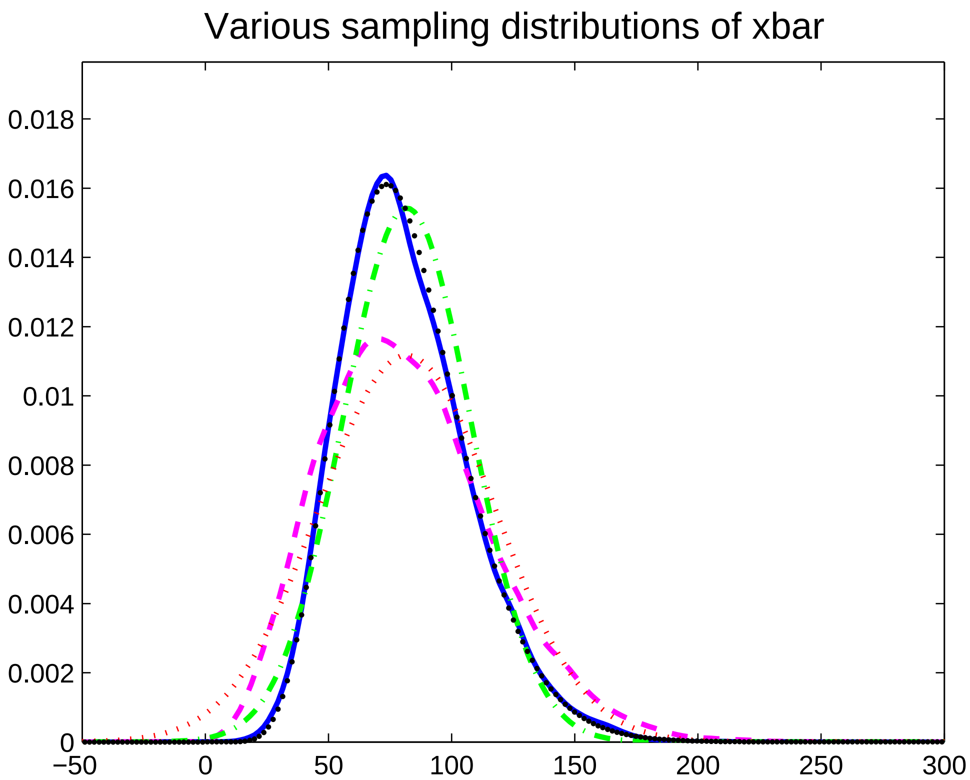

Figure 1.6: *Various estimates of the sampling distributions of* $\bar{X}$*. Dotted (big dots) is the assumed parametric density (gamma), Dash-dot is the parametric normal approximation, Solid is the parametric bootstrap approximation, Dotted (small dots) is the nonparametric normal appoximation and 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 te 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:
   $\widehat{E}[\bar{X}] = \bar{x} = 81.80$ and $\widehat{Std}[\bar{X}] = \bar{x}/\sqrt{n} = 25.87$
2. asymptotic parametric:
   $\widehat{E}[\bar{X}] = \bar{x} = 81.80$ and $\widehat{Std}[\bar{X}] = \bar{x}/\sqrt{n} = 25.87$
3. bootstrap parametric:
   $\widehat{E}[\bar{X}] \approx 82.15$ and $\widehat{Std}[\bar{X}] \approx 26.04$
4. asymptotic Non-parametric:
   $\widehat{E}[\bar{X}] = \bar{x} = 81.80$ and $\widehat{Std}[\bar{X}] = s/\sqrt{n} = 35.71$
5. bootstrap Non-parametric:
   $\widehat{E}[\bar{X}] \approx 82.52$ and $\widehat{Std}[\bar{X}] \approx 33.74$

• A Simulated Example

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

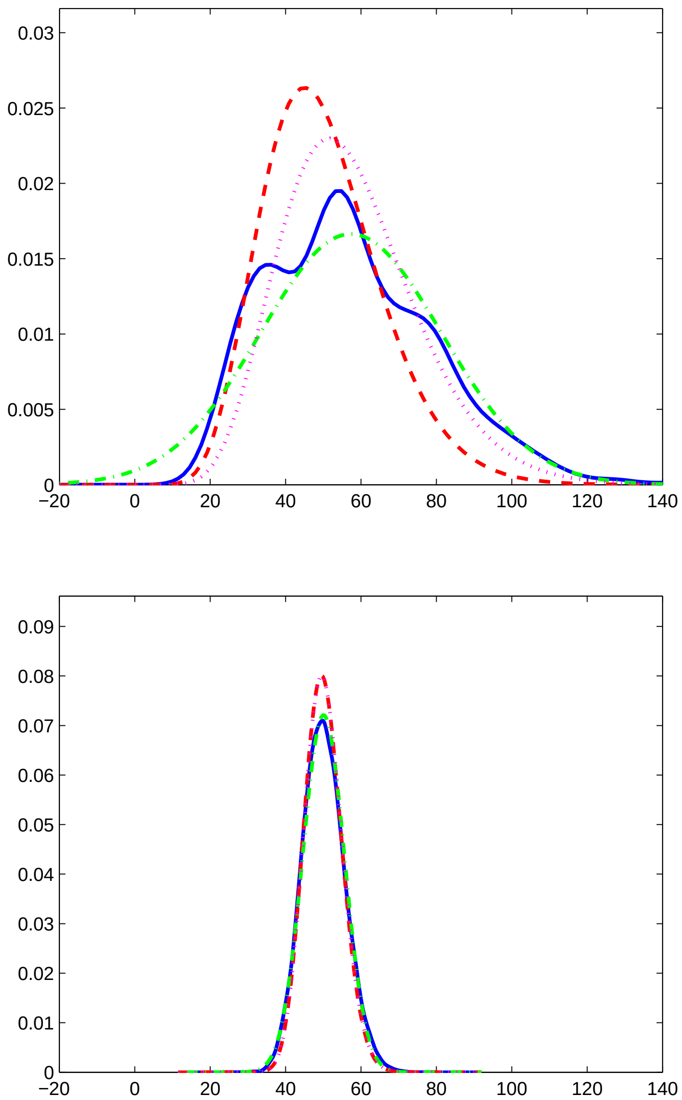

Figure 1.7: *Sampling distribution of* $\bar{X}$ *where* $X_i \sim$ *Expo* $(\mu = 50)$. *Dashed line is the true Gamma* $\Gamma(n, \mu/n)$, *dot is the estimated Gamma* $\Gamma(n, \bar{X}/n)$, *dash-dot is the estimated Normal approximation* $N(\bar{X}, s^2/n)$ *and solid is the Nonparametric Bootstrap 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 $T = T(\mathcal{X})$?
  - 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 $T = T(\mathcal{X})$ can be represented in terms of simpler statistics $U_j = U_j(\mathcal{X}), j = 1, \ldots, p$, such as sample moments, for which asymptotic sampling distribution can be derived by using the CLT.

Theorem 1.1. Delta Method. *Let $\sqrt{n}(U - \mu) \sim AN_p(0, \Sigma)$ where $\Sigma > 0$ and $T = g(U)$ where $g : I\!R^p \to I\!R^q$, $q \leq p$ and g are differentiable at $u = \mu$. Then*

$$\sqrt{n}(T - g(\mu)) \sim AN_q(0, \Delta' \, \Sigma \, \Delta)$$

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 $\sigma^2 = \Sigma = n\,Var(U)$.

– 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_θ 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:

$$\begin{aligned} L_\theta(x; F) &= \lim_{\varepsilon \to 0} \frac{\theta\left((1-\varepsilon)F + \varepsilon H_x\right) - \theta(F)}{\varepsilon} \\ &= \left. \frac{\partial \theta\left((1-\varepsilon)F + \varepsilon H_x\right)}{\partial \varepsilon} \right|_{\varepsilon = 0}. \end{aligned}$$

- $L_\theta(x; F)$ is called the **influence function** of θ 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 $G = F_n$:

$$\begin{aligned}\theta(F_n) &= \theta(F) + \int L_\theta(x;F)\,dF_n(x) + o_p(||F_n - F||)\\ &\approx \theta(F) + \int L_\theta(x;F)\,dF_n(x)\\ &= \theta(F) + \frac{1}{n}\sum_{i=1}^{n} L_\theta(x_i;F).\end{aligned}$$

– 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:* $\theta(F) = \int x dF(x)$ and $\theta(F_n) = \bar{X}$. We have

- $\theta\left((1-\varepsilon)F + \varepsilon H_x\right) = (1-\varepsilon)\mu + \varepsilon x$ so that

$$L_\theta(x; F) = \left.\frac{\partial\left((1-\varepsilon)\mu + \varepsilon x\right)}{\partial \varepsilon}\right|_{\varepsilon=0} = x - \mu.$$

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

- Finally:

$$var_L(F_n) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n^2}s^2 = \frac{v}{n}.$$

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

- So that the nonparametric delta method provides

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

– *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 $\mu_{rs} = E[X^r Y^s]$. 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 Jacknife: 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

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

    Equivalently:

$$Prob_{F_n}\left\{|Bias^*_B(T) - Bias^*_\infty(T)| \leq 2\,\frac{Std^*(T)}{\sqrt{B}}\right\} \approx 0.95$$

The "Error Bound" $2\,Std^*(T)/\sqrt{B}$ 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
```

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

Table 1.1: *Choice of B for Bias estimation of* μ_X/μ_Y.

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 $AStd(\hat{\theta}) = 1/\sqrt{8} = 0.3536$.

Value of B	$Bias^*_B(T)$	$Std^*_B(T)$	error bound
100	-0.0680	0.4069	0.0814
200	-0.0751	0.4052	0.0573
400	-0.0814	0.3984	0.0398
800	-0.0949	0.4310	0.0305
1000	-0.0543	0.4050	0.0256
2000	-0.0702	0.4075	0.0182
5000	-0.0750	0.4119	0.0117
10000	-0.0762	0.4140	0.0083

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

1.4.2 Correction for the bias

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

- Usually

$$\begin{aligned} T_{corr}(\mathcal{X}) &= T(\mathcal{X}) - Bias^*(T) \\ &= 2\,T(\mathcal{X}) - \frac{1}{B}\sum_{b=1}^{B} T(\mathcal{X}^{*(b)}) \end{aligned}$$

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

– How to compute MSE_F?

- we have

$$
\begin{aligned}
Var_F(T_{corr}(\mathcal{X})) &= Var_F\left(2\,T(\mathcal{X}) - \frac{1}{B}\sum_{b=1}^{B} T(\mathcal{X}^{*(b)})\right)\\
&= 4\,Var_F(T(\mathcal{X})) + Var_F\left(\frac{1}{B}\sum_{b=1}^{B} T(\mathcal{X}^{*(b)})\right)\\
&\quad -2\,Cov_F\left(2\,T(\mathcal{X}), \frac{1}{B}\sum_{b=1}^{B} T(\mathcal{X}^{*(b)})\right)\\
&= 4\,Var_F(T(\mathcal{X})) + Var_F[E^*(T(\mathcal{X}^*))]\\
&\quad -4\,Cov_F[(T(\mathcal{X}), E^*(T(\mathcal{X}^*)] + O(B^{-1})
\end{aligned}
$$

which contains many unknown quantities.

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

– 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:

$$\text{If } \frac{|Bias^*(T)|}{Std_B^*(T)} \leq 0.25 \text{ don't correct for the bias,}$$

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

$$\begin{aligned} RMSE_F(T) &= Std_F(T)\sqrt{1 + \left(\frac{Bias_F(T)}{Std_F(T)}\right)^2} \\ &\approx Std_F(T)\left\{1 + \frac{1}{2}\left(\frac{Bias_F(T)}{Std_F(T)}\right)^2\right\} \end{aligned}$$

– In the examples above (with $B = 5000$):

- For the estimation of the ratio of means we have

$$|Bias^*(T)|/Std_B^*(T) = 0.6247/2.0510 = 0.3046$$
$$\hat{\theta}_{corr} = \hat{\theta} - Bias^*(T) = 2.8574 - 0.6247 = 2.2327$$

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

$$|Bias^*(T)|/Std_B^*(T) = 0.0750/0.4119 = 0.1821$$
$$\hat{\theta}_{corr} = \hat{\theta} - Bias^*(T) = 3.8368 - (-0.0750) = 3.9118$$

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

$$BIAS(\hat{\theta}) \approx \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta_0)$$
$$MSE(\hat{\theta}) \approx \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_m - \theta_0)^2$$
$$BIAS(\hat{\theta}_{corr}) \approx \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_{corr,m} - \theta_0)$$
$$MSE(\hat{\theta}_{corr}) \approx \frac{1}{M}\sum_{m=1}^{M}(\hat{\theta}_{corr,m} - \theta_0)^2$$

  – Example 1: simlation 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), $\theta_0 = \mu_X/\mu_Y = 2$

$$BIAS(\hat{\theta}) = 0.2175, \quad MSE(\hat{\theta}) = 1.2558$$
$$BIAS(\hat{\theta}_{corr}) = -0.0014, \quad MSE(\hat{\theta}_{corr}) = 0.9871.$$

– 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), $\theta_0 = \log(\mu) = 4.6052$:

$$BIAS(\hat{\theta}) = -0.0631, \quad MSE(\hat{\theta}) = 0.1377$$
$$BIAS(\hat{\theta}_{corr}) = -0.0099, \quad MSE(\hat{\theta}_{corr}) = 0.1358.$$

```
% MATLAB code
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)

BIASt=mean(MC_that)-ttrue;
BIAStcorr=mean(MC_tcorr)-ttrue;
MSEt=mean((MC_that-ttrue*ones(MC,1)).^2);
MSEtcorr=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 $T(\mathcal{X}) - \theta(F)$:

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

– We have:

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

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

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

– 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):

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

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

$$\mathcal{F}_n(\theta) = Var(\dot{\ell}(\mathcal{X};\theta)) = -E\left[\ddot{\ell}(\mathcal{X};\theta)\right].$$

- $\mathcal{F}_n(\theta)$ can be estimated by $\mathcal{F}_n(\hat{\theta})$, or more simply by its consistent estimator, the observed Fisher information $-\ddot{\ell}(\mathcal{X};\hat{\theta})$.

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

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

where $\widehat{\sigma}^2 = AVar(T(\mathcal{X}))$.

– 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 Appraoch": We can also use the (eventual) asymptotic normality, but the variance of $T(\mathcal{X})$, is obtained by the bootstrap approximation:

[1] provide the boostrap estimate of the sampling distribution of $T(\mathcal{X})$: $\widehat{G}(x) = Dist^*_T(x)$;

[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 $\theta(F)$ 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 $\widehat{G}(x)$ 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:

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

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

$$\begin{aligned}\widehat{H}(x) &= P_{F_n}\left(T(\mathcal{X}^*) - \theta(F_n) \leq x\right) \\ &= P\left(T(\mathcal{X}^*) - \theta(F_n) \leq x \,|\, \mathcal{X}\right) \\ &\approx \frac{1}{B}\sum_{b=1}^{B} I\left(T(\mathcal{X}^{*(b)}) - \theta(F_n) \leq x\right).\end{aligned}$$

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

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

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

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

$$P_{F_n}\Big(T(\mathcal{X}^*) \leq v^*(a)\Big) = \widehat{G}(v^*(a)) = a$$

– Since $\widehat{H}(x) = \widehat{G}(x + \theta(F_n))$ we just have to shift the obtained quantiles to get $u^*(a)$:

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

• The quantiles $v^*(a)$ of $\widehat{G}(x) = Dist^*_T(x)$ 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 $k1 = [B\ \frac{\alpha}{2}] + 1$ and $k2 = [B\ (1 - \frac{\alpha}{2})]$, where $[\cdot]$ stands for the integer part.

- Finally:

$$v^*(\frac{\alpha}{2}) = T^*_{(k1)}$$
$$v^*(1 - \frac{\alpha}{2}) = T^*_{(k2)}$$

- 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:

$$\Big[T(\mathcal{X}) + \theta(F_n) - v^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) + \theta(F_n) - v^*(\frac{\alpha}{2})\Big]$$
$$= \Big[2\,T(\mathcal{X}) - v^*(1 - \frac{\alpha}{2}), 2\,T(\mathcal{X}) - v^*(\frac{\alpha}{2})\Big]$$

- **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 $s = 112.94$.

1. Parametric model: If X is exponential (?) we have an exact confidence interval and a pivotal root (available in this simple case).
$$\frac{\bar{X}}{\mu} \sim \text{Gamma } (n, \frac{1}{n})$$
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 Gamma$(n, 1/n)$.

2. Asymptotic parametric approximation: If X is exponential (?) we have an asymptotical pivotal root.
$$\frac{\bar{X} - \mu}{\mu/\sqrt{n}} \sim AN(0, 1)$$
and by Slutzky's theorem $\sqrt{n}(\bar{X} - \mu)/\bar{X} \sim AN(0, 1)$. 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 $\sqrt{n}(\bar{X} - \mu)/S \sim AN(0, 1)$. 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 $v^*(a) = \widehat{G}^{-1}(a)$ 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 $\mathcal{X}$.

   (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 $\Gamma(n, 1/n)$.

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

   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.

method	lower limit	upper limit
Parametric Gamma	47.8788	170.5805
Parametric Pivotal-bootstrap	47.7969	168.9480
Parametric Bootstrap	23.6067	123.9947
Parametric Asymp. Normal	31.1008	132.4992
Nonparametric Asymp. Normal	11.8014	151.7986
Nonparametric Bootstrap	6.8000	137.3000

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 σ^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 ($B = 5000$).

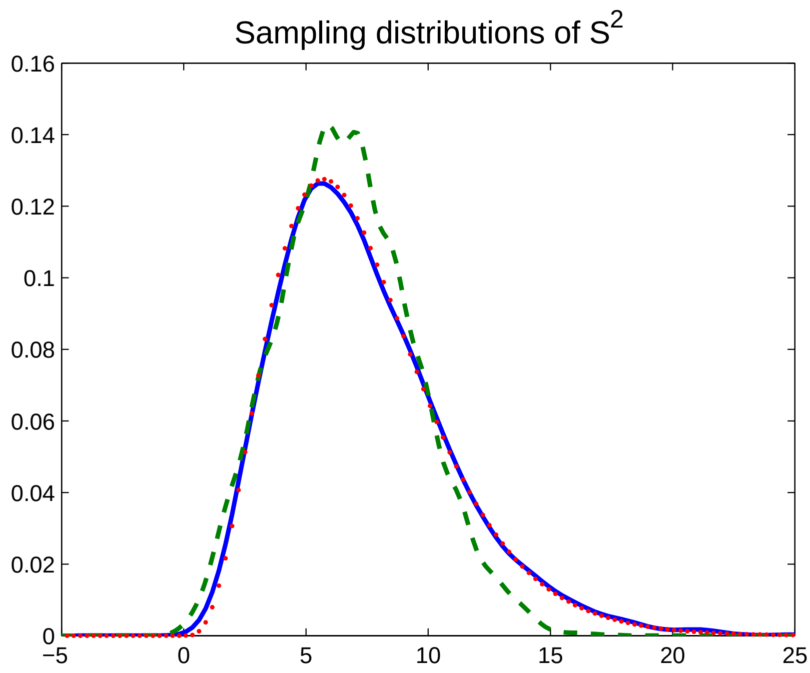

Figure 1.8: *Estimates of the sampling distribution of S^2: estimated Gamma (dotted), parametric bootstrap (solid), nonparametric bootstrap (dashed).*

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

$$Bias^*(S^2) = 0.0575 \text{ and } Std^*(S^2) = 2.7492$$
$$Bias^*(V) = -0.6094 \text{ and } Std^*(V) = 2.4743$$

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

The confidence intervals for σ^2 obtained by the parametric and non-parametric bootstrap distribution of S^2 are:

$$\text{parametric bootstrap} = [-1.7756\ ,\ 11.7803]$$
$$\text{nonparametric bootstrap} = [1.3987\ ,\ 12.2273]$$

1.6 Summary Table of the Bootstrap Ideas

REAL WORLD	BOOTSTRAP WORLD
F unknown $\theta(F)$ unknown $\boxed{\text{DGP } F \ : \ X \sim F}$ $\Downarrow$ $\mathcal{X} = (X_1, \ldots, X_n)$	Given $\mathcal{X} = (X_1, \ldots, X_n)$, F_n is known $\theta(F_n)$ known $\boxed{\text{DGP}^* \ F_n \ : \ X^* \sim F_n}$ $\Downarrow$ $\mathcal{X}^* = (X_1^*, \ldots, X_n^*)$
$T(\mathcal{X})$ estimator of $\theta(F)$ $G(x) = P_F(T(\mathcal{X}) \leq x)$ Moments $E_F(T(\mathcal{X}))$ $Var_F(T(\mathcal{X}))$	$T(\mathcal{X}^*)$ estimator of $\theta(F_n)$ $\widehat{G}(x) = P_{F_n}(T(\mathcal{X}^*) \leq x)$ Moments $E_{F_n}(T(\mathcal{X}^*))$ $Var_{F_n}(T(\mathcal{X}^*))$
$W = T(\mathcal{X}) - \theta(F)$ $H(x) = P_F(W \leq x) \Rightarrow u(a) = H^{-1}(a)$ $\Downarrow$ CI for θ $[T(\mathcal{X}) - u(1-\alpha/2), T(\mathcal{X}) - u(\alpha/2)]$	$W^* = T(\mathcal{X}^*) - \theta(F_n)$ $\widehat{H}(x) = P_{F_n}(W^* \leq x) \Rightarrow \hat{u}(a) = \widehat{H}^{-1}(a)$ $\Downarrow$ Bootstrap CI for θ $[T(\mathcal{X}) - \hat{u}(1-\alpha/2), T(\mathcal{X}) - \hat{u}(\alpha/2)]$
$S = \frac{T(\mathcal{X})-\theta(F)}{\hat{\sigma}(\mathcal{X})/\sqrt{n}}$ $K(x) = P_F(S \leq x) \Rightarrow y(a) = K^{-1}(a)$ $\Downarrow$ CI for θ $\left[T(\mathcal{X}) - y(1-\alpha/2)\frac{\hat{\sigma}(\mathcal{X})}{\sqrt{n}}, T(\mathcal{X}) - y(\alpha/2)\frac{\hat{\sigma}(\mathcal{X})}{\sqrt{n}}\right]$	$S^* = \frac{T(\mathcal{X}^*)-\theta(F_n)}{\hat{\sigma}(\mathcal{X}^*)/\sqrt{n}}$ $\widehat{K}(x) = P_{F_n}(S^* \leq x) \Rightarrow \hat{y}(a) = \widehat{K}^{-1}(a)$ $\Downarrow$ Bootstrap CI for θ $\left[T(\mathcal{X}) - \hat{y}(1-\alpha/2)\frac{\hat{\sigma}(\mathcal{X})}{\sqrt{n}}, T(\mathcal{X}) - \hat{y}(\alpha/2)\frac{\hat{\sigma}(\mathcal{X})}{\sqrt{n}}\right]$

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 η as

$$[U(\mathcal{X}) - u_U(1-\alpha/2), U(\mathcal{X}) - u_U(\alpha/2)],$$

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

$$[U(\mathcal{X}) + u_U(\alpha/2), U(\mathcal{X}) + u_U(1-\alpha/2)]$$

- Basic Bootstrap approximation:
  - If we knew h, we could use the basic bootstrap method to obtain an approximation of the confidence interval for η 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 $U^* - \widehat{\eta}$, with $U^* = U(\mathcal{X}^*) = h(T(\mathcal{X}^*))$ and $\widehat{\eta} = U(\mathcal{X}) = h(T(\mathcal{X}))$.
  - 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 η can be transformed back to a confidence interval for θ,
  - 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 precentile 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 θ scale:

$$\text{Percentile} \quad \Big[T(\mathcal{X}) + u^*(\frac{\alpha}{2}), T(\mathcal{X}) + u^*(1-\frac{\alpha}{2})\Big]$$
$$\text{Basic bootstrap} \quad \Big[T(\mathcal{X}) - u^*(1-\frac{\alpha}{2}), T(\mathcal{X}) - u^*(\frac{\alpha}{2})\Big]$$

• 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:

$$\text{Percentile} \quad \Big[v^*(\frac{\alpha}{2}), v^*(1-\frac{\alpha}{2})\Big]$$
$$\text{Basic bootstrap} \quad \Big[2\,T(\mathcal{X}) - v^*(1-\frac{\alpha}{2}), 2\,T(\mathcal{X}) - v^*(\frac{\alpha}{2})\Big]\,.$$

• **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 η.

– 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 $U(\mathcal{X})$:

$$\begin{aligned} h(T) &\sim N(h(\theta) - w, 1) \\ U - \eta &\sim N(-w, 1) \end{aligned}$$

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

• How to estimate w ?

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

$$\begin{aligned} P(U > \eta) &= P(Z - w > 0) \\ &= P(Z > w). \end{aligned}$$

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

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

– This involve an equation in w:

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

the solution is:

$$w = \Phi^{-1}\left(P(T \leq \theta)\right).$$

– The bootstrap estimate of w is then:

$$\begin{aligned}
\widehat{w} &= \Phi^{-1}\left(P(T(\mathcal{X}^*) \leq \theta(F_n) \,|\, \mathcal{X})\right) \\
&= \Phi^{-1}\left(P_{F_n}(T(\mathcal{X}^*) \leq \theta(F_n))\right) \\
&= \Phi^{-1}\left(\hat{G}(\theta(F_n))\right) \\
&= \Phi^{-1}\left(\frac{\#\left[T(\mathcal{X}^{*(b)}) \leq \theta(F_n)\right]}{B}\right)
\end{aligned}$$

- How to obtain confidence interval for θ?

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

$$\begin{aligned}
&P\left(z_{\alpha/2} \leq U - \eta + w \leq z_{1-\alpha/2}\right) = 1 - \alpha \\
&P\left(U + w - z_{1-\alpha/2} \leq \eta \leq U + w - z_{\alpha/2}\right) = 1 - \alpha \\
&P\left(U + w + z_{\alpha/2} \leq \eta \leq U + w + z_{1-\alpha/2}\right) = 1 - \alpha \\
&P\left(\hat{\eta}_{\alpha/2} \leq \eta \leq \hat{\eta}_{1-\alpha/2}\right) = 1 - \alpha,
\end{aligned}$$

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

– Finally, confidence interval for θ: 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

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

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

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

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

$$\widehat{G}(\hat{\theta}_a) \approx \Phi\left(2\,w + z_a\right)$$

- Finally we obtain

$$\begin{aligned}\hat{\theta}_a &= \widehat{G}^{-1}\left(\Phi(2\,w + z_a)\right) \\ &= v^*\left(\Phi(2\,w + z_a)\right)\end{aligned}$$

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

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

- 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 unknwon 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 η:

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

  where $\sigma_\eta = \sigma_{\eta_0}[1+a(\eta-\eta_0)]$ where η_0 is any convenient reference point on the scale of η values: the idea is that we approximate (when n is large) σ_η by a linear function. To simplify we will chose below η_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 monotome function h.

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

  – 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 η**

  – From H_{BC_a}, 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 η is thus obtained by

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

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

- **Confidence intervals for θ**

We have to come back to the θ 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 $\widehat{G}(\cdot)$ be the cdf of $\hat{\theta}^* = T(\mathcal{X}^*)$. We have, in the bootstrap world:

$$\begin{aligned}\widehat{G}(\hat{\theta}_\beta) &= \text{Prob}(\hat{\theta}^* \leq \hat{\theta}_\beta|\mathcal{X}) = \text{Prob}(\hat{\eta}^* \leq \hat{\eta}_\beta|\mathcal{X}) \\ &= \text{Prob}\left(\hat{\eta}^* \leq \hat{\eta} + \sigma_{\hat{\eta}}\frac{w+z_\beta}{1-a(w+z_\beta)}|\mathcal{X}\right) \\ &= \text{Prob}\left(\frac{\hat{\eta}^*-\hat{\eta}}{\sigma_{\hat{\eta}}} + w \leq w + \frac{w+z_\beta}{1-a(w+z_\beta)}|\mathcal{X}\right).\end{aligned}$$

– 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:

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

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

$$\begin{aligned}\hat{\theta}_\beta &= \widehat{G}^{-1}\left[\Phi\left(w + \frac{w+z_\beta}{1-a(w+z_\beta)}\right)\right] \\ &= v^*\left[\Phi\left(w + \frac{w+z_\beta}{1-a(w+z_\beta)}\right)\right],\end{aligned}$$

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 quanriles 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:

$$\begin{aligned} G(\theta) &= \text{Prob}_F(\hat{\theta} \leq \theta) = \text{Prob}_F(\hat{\eta} \leq \eta) \\ &= \text{Prob}_F(\frac{\hat{\eta} - \eta}{\sigma_\eta} \leq 0) = \text{Prob}(Z \leq w) = \Phi(w) \end{aligned}$$

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

$$\begin{aligned} \hat{w} &= \Phi^{-1}\big(\text{Prob}(\hat{\theta}^* \leq \hat{\theta}|\mathcal{X})\big) = \Phi^{-1}(\widehat{G}(\hat{\theta})) \\ &= \Phi^{-1}\left(\frac{\#\{\hat{\theta}^{*,b} \leq \hat{\theta}\}}{B}\right). \end{aligned}$$

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

$$\hat{a} = \frac{\sum_{i=1}^n L_\theta^3(x_i, F_n)}{6\big[\sum_{i=1}^n L_\theta^2(x_i, F_n)\big]^{3/2}},$$

  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 ith jacknife value:

$$\hat{a} = \frac{\sum_{i=1}^n \big(\hat{\theta}_{(\bullet)} - \hat{\theta}_{(i)}\big)^3}{6\Big[\sum_{i=1}^n \Big(\hat{\theta}_{(\bullet)} - \hat{\theta}_{(i)}\Big)^2\Big]^{3/2}},$$

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

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

where $\sigma^2(F) = n\,Var_F(T(\mathcal{X}))$.

– 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 σ is **known**: confidence interval for $\theta(F)$:

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

• If σ is **unknown**: "<u>Studentization</u>"

– 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 $\widehat{H}$:

$$\begin{aligned}\widehat{H}(x) &= P(S^* \leq x \,|\, \mathcal{X}) \\ &= P\left(\sqrt{n}\left(\frac{T(\mathcal{X}^*) - \theta(F_n))}{\sigma(F_n)}\right) \leq x \mid \mathcal{X}\right).\end{aligned}$$

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

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

- Bootstrap confidence intervals:

  – **If σ is known**:

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

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

– **If σ 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 $\widehat{K}(x)$.

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

  – An approximate end point of level α, say $\hat{\theta}_\alpha$ is such that 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 $\widehat{K}(x)$, 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.

method	lower limit	upper limit
Parametric Gamma	47.8788	170.5805
Parametric Pivotal-bootstrap	47.7969	168.9480
Parametric Bootstrap	23.6067	123.9947
Parametric Asymp. Normal	31.1008	132.4992
Nonparametric Asymp. Normal	11.8014	151.7986
Nonparametric Basic Bootstrap	6.8000	137.3000
Nonparametric Percentile	26.2500	156.8000
Nonparametric BC-Percentile	28.7000	164.2567
Nonparametric BC_a-Percentile	35.3754	184.9024
Nonparametric Bootstrap-t	23.3073	230.1239

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

2.5 Some Simulated Examples

- **Example 1:** $X_i \sim$ Expo (μ), 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 μ at the level 95%. For the bootstrap $B = 5000$. Results in Table 2.2.

method	lower limit	upper limit
Estimated Gamma	4.5489	16.2066
Estimated As. Normal	2.9548	12.5885
Basic Bootstrap	5.1181	10.1953
Percentile Bootstrap	5.3481	10.4253

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:

$$\text{Error bound } = 2\sqrt{\frac{0.05 * 0.95}{1000}} = 0.0138$$

    - Table 2.3 displays some values of the error bound as a function of M.

Value of M	Error bound
100	0.0436
200	0.0308
500	0.0195
1000	0.0138
2000	0.0097
5000	0.0062
10000	0.0044

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 $\bar{X}$.

sample size	$n = 5$		$n = 10$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
Exact Gamma	26.25	0.944	14.90	0.958	3.97	0.949	1.77	0.959
As. Normal	17.76	0.874	12.32	0.908	3.90	0.933	1.76	0.963
Basic Boot.	13.39	0.765	10.68	0.839	3.83	0.924	1.75	0.961
Perc. Boot.	13.39	0.782	10.68	0.865	3.83	0.933	1.75	0.958

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 $X \sim Expo(\mu)$: 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, $N \sim Poisson(1/\mu)$

- Parameter of interest

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

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

- What are the sampling properties of $\hat{\pi}$? We have here (delta method) an asymptotic result:

$$\hat{\pi} \sim AN\left(\pi, \frac{e^{-2/\mu}}{n\mu^2} = \frac{1}{n}\left(\frac{\pi}{\mu}\right)^2\right)$$

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

– 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 $n = 100$, $\hat{\pi}_{obs} = 0.0.9136$.

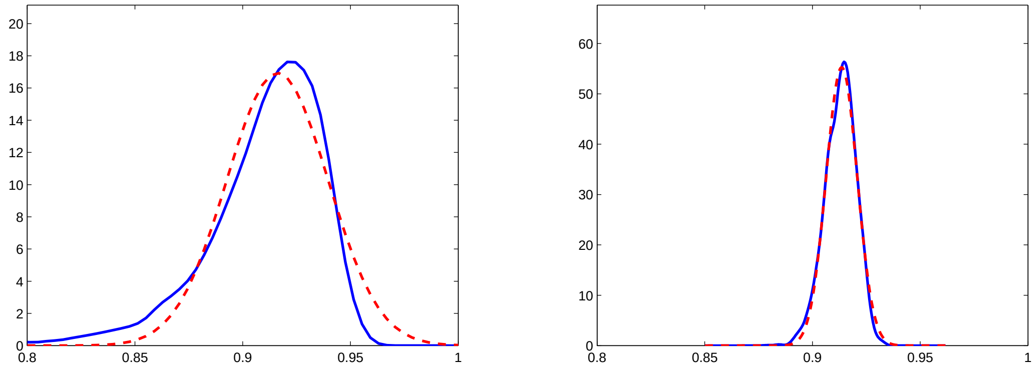

Figure 2.1: *Bootstrap approximation for the sampling distribution of* $\hat{\pi} = \exp(-1/\bar{X})$*: here* $\pi = 0.9048$ *for* $n = 10$ *(left)* $\hat{\pi}_{obs} = 0.9161$ *and for* $n = 100$ *(right),* $\hat{\pi}_{obs} = 0.9136$*.Here* $B = 5000$*. Dashed line is the estimated Normal approximation and solid is the Bootstrap approximation*

– Monte-Carlo evaluation of the performances of confidence intervals for π.

- The asymtotic CI for π 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 bootstap-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

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

- The results are in Table 2.5.

sample size	$n = 5$		$n = 10$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	0.186	0.956	0.122	0.957	0.036	0.952	0.016	0.943
Basic Boot.	0.217	0.817	0.130	0.879	0.036	0.938	0.016	0.941
Perc. Boot.	0.217	0.801	0.130	0.874	0.036	0.942	0.016	0.941
BC-perc.	0.198	0.808	0.124	0.883	0.036	0.940	0.016	0.939
Bootstrap-t	0.155	0.857	0.107	0.908	0.035	0.945	0.016	0.938

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

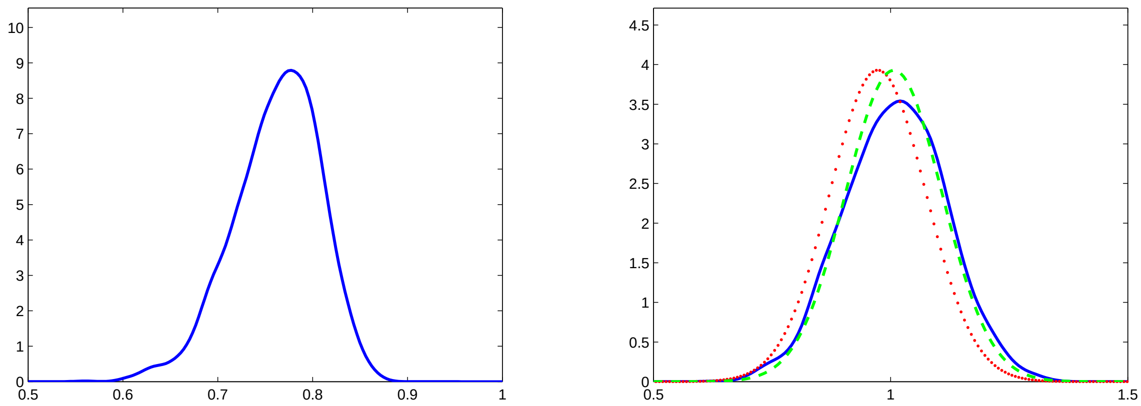

Figure 2.2: *Bootstrap approximation for the sampling distribution of* $\hat{\rho} = r$ *(left) and of* $\hat{z} = \tanh^{-1}(r)$ *(right), original sample has* $n = 100$ *and* $r = 0.7644$ *whereas the true* $\rho = 0.75$*. Dot is the true asymptotic normal (with* $\rho = 0.75$*), dashed is the estimated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

– Monte-Carlo evaluation of the performances of CI for ρ. The asymtotic CI for ρ 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.

sample size	$n = 10$		$n = 20$		$n = 100$	
method	length	coverage	length	coverage	length	coverage
As. Normal	0.68	0.953	0.43	0.944	0.18	0.955
Basic Boot.	0.65	0.778	0.40	0.830	0.17	0.912
Perc. Boot.	0.65	0.922	0.40	0.908	0.17	0.945
BC-Perc.	0.73	0.944	0.42	0.926	0.17	0.943

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:

$$\begin{aligned} Cov(X_1, X_2) &= E(X_1 X_2) - E(X_1)E(X_2) \\ &= E(X_1^3) - 0 = 0 \end{aligned}$$

    - $\rho = 0$ but X_1 and X_2 are not independent!
  - Monte-Carlo evaluation of different confidence intervals see Table 2.7

sample size	$n = 10$		$n = 50$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	0.95	0.628	0.51	0.616	0.38	0.628	0.17	0.635
Basic Boot.	1.49	0.431	0.95	0.722	0.74	0.816	0.37	0.916
Perc. Boot.	1.49	0.847	0.95	0.891	0.74	0.915	0.37	0.947
BC-Perc.	1.58	0.893	0.95	0.898	0.74	0.918	0.37	0.948

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

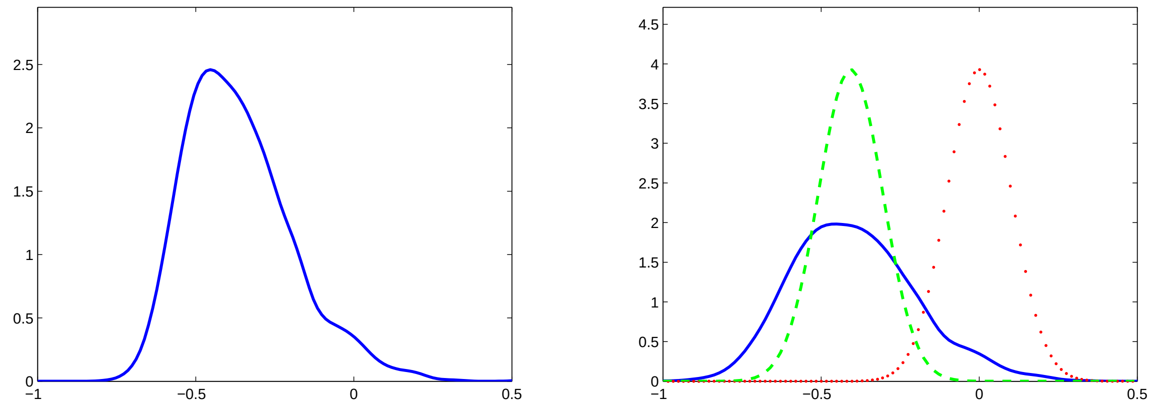

Figure 2.3: *Bootstrap approximation for the sampling distribution of* $\hat{\rho} = r$ *(left) and of* $\hat{z} = \tanh^{-1}(r)$ *(right), original sample has* $n = 100$ *and* $r = 0.0942$ *whereas the true* $\rho = 0$. *Dot is the "true" but wrong asymptotic normal (with* $\rho = 0$*), dashed is the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

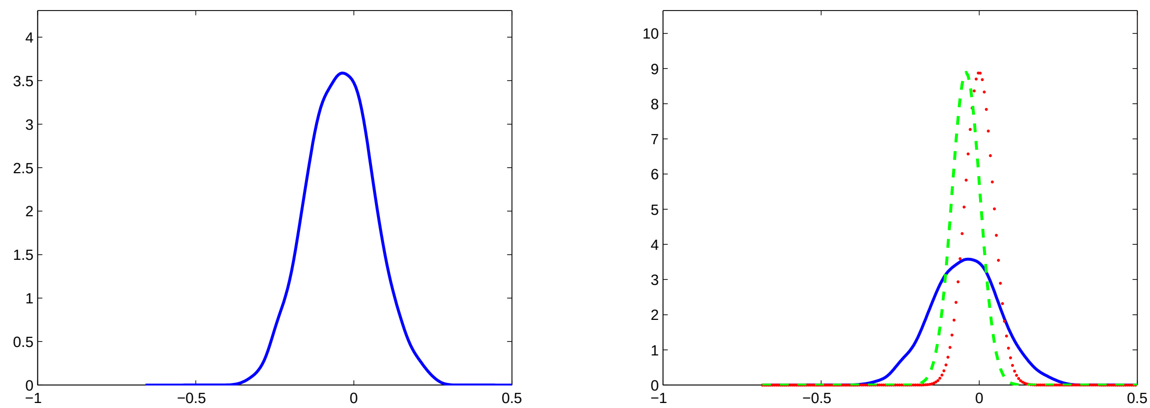

Figure 2.4: *Bootstrap approximation for the sampling distribution of* $\hat{\rho} = r$ *(left) and of* $\hat{z} = \tanh^{-1}(r)$ *(right), original sample has* $n = 500$ *and* $r = -0.1183$ *whereas the true* $\rho = 0$. *Ddot is the "true" but wrong asymptotic normal (with* $\rho = 0$*), dashed is 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 θ.

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

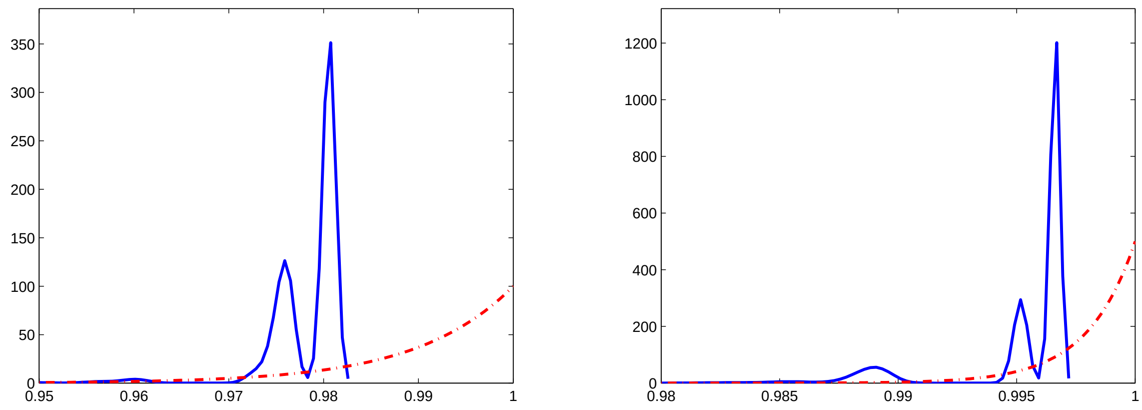

Figure 2.5: *Bootstrap approximation for the sampling distribution of $X_{(n)}$, original sample has $n = 100$ and $x_{(n)} = 0.9806$ (left) and $n = 500$ and $x_{(n)} = 0.9966$ (right). Here $B = 2000$. Solid line is the bootstrap approximation, dashdot line is the true sampling density.*

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.

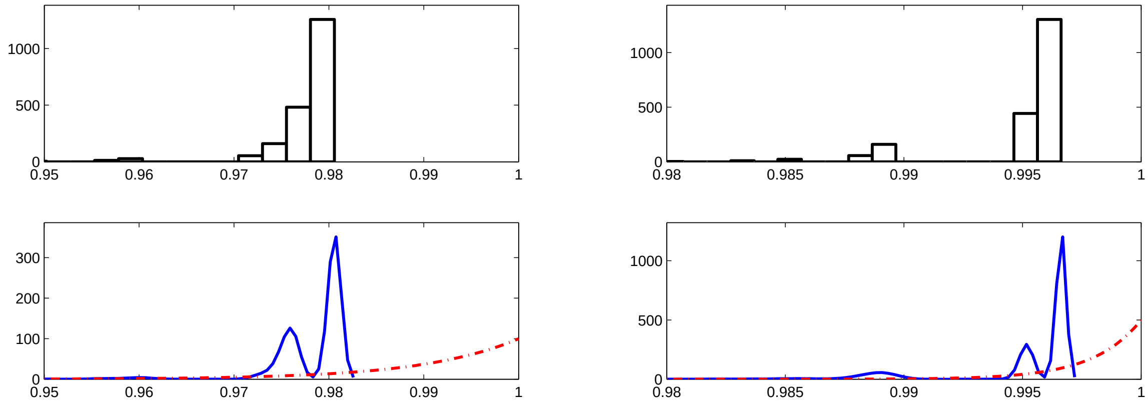

Figure 2.6:

But there are solutions: smoothed bootstrap and subsampling.

– *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 symetric distribution for $X_{(n)}$.

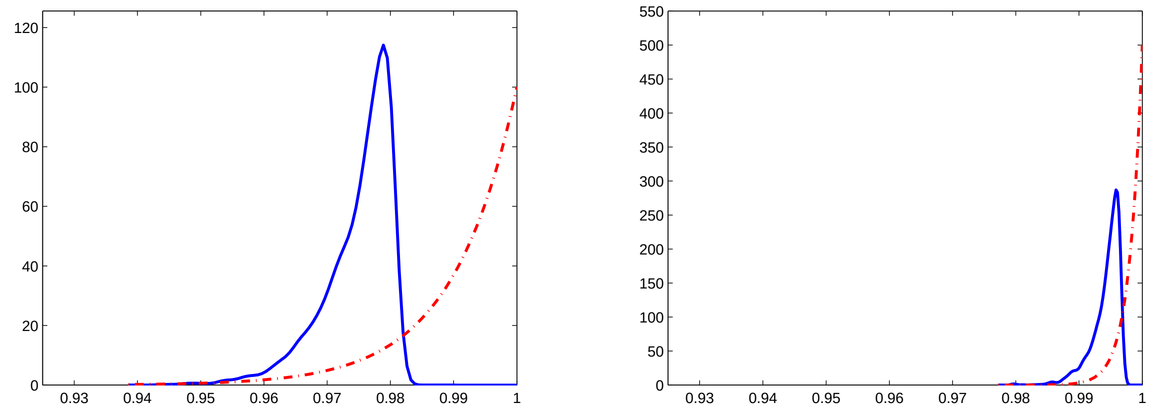

Figure 2.7: *Smoothed-Bootstrap approximation of the sampling distribution of* $X_{(n)}$*, original sample has* $n = 100$ *and* $x_{(n)} = 0.9806$ *(left) and* $n = 500$ *and* $x_{(n)} = 0.9966$ *(right). Here* $B = 2000$*. Solid line is the bootstrap approximation, dashdot line is the true sampling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

sample size	$n = 10$		$n = 50$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
Basic Boot.	0.2865	0.8790	0.0748	0.9340	0.0386	0.9350	0.0077	0.9450
Perc. Boot.	0.2865	0.5480	0.0748	0.6140	0.0386	0.6310	0.0077	0.6430
BC-Perc.	0.3895	0.5150	0.1077	0.5890	0.0559	0.5980	0.0112	0.6110

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 $\mathcal{X}$ a random sample of size $m = [n^{\gamma}]$ where $\gamma < 1$.

  – Not easy to choose γ: 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 $Expo(\mu)$

```
% 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 $X \sim Expo(\mu)$

```
% 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);
lowperc=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;lowperc 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))
avlengtboot=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

```
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, bandwith 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 $\sigma^2(F) = nAVar_F(T(\mathcal{X}))$.

– 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 σ 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:

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

First Order Accuracy of the asymptotic approximation

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

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

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

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

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

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:

$$Dist^*_S(x) - Dist_{S,F}(x) = O_p(n^{-1})$$

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 σ 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:

$$W = \sqrt{n}(T(\mathcal{X}) - \theta(F))$$
$$W^* = \sqrt{n}(T(\mathcal{X}^*) - T(\mathcal{X}))$$

  – 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:

$$\begin{aligned} Dist_{W,F}(x) &= P_F(W \leq x) = P_F(S \leq \frac{x}{\sigma}) \\ &= \Phi(\frac{x}{\sigma}) + n^{-\frac{1}{2}}p(\frac{x}{\sigma})\phi(\frac{x}{\sigma}) + O(n^{-1}) \end{aligned}$$

$$\begin{aligned} Dist^*_W(x) &= P_{F_n}(W^* \leq x) = P_{F_n}(S^* \leq \frac{x}{\hat{\sigma}}) \\ &= \Phi(\frac{x}{\hat{\sigma}}) + n^{-\frac{1}{2}}\hat{p}(\frac{x}{\hat{\sigma}})\phi(\frac{x}{\hat{\sigma}}) + O_p(n^{-1}) \end{aligned}$$

  – Now we have:

$$\begin{aligned} Dist^*_W(x) - Dist_{W,F}(x) &= \Phi(\frac{x}{\hat{\sigma}}) - \Phi(\frac{x}{\sigma}) + O_p(n^{-1}) \\ &= (\hat{\sigma} - \sigma)R(x) + O_p(n^{-1}) \end{aligned}$$

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

$$Dist^*_W(x) - Dist_{W,F}(x) = O_p(n^{-\frac{1}{2}})$$

  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 σ is unknown and no estimators available.

– $Dist^*_W(x)$ 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 $Dist^*_W(x)$.

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

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

• If such an estimator of σ 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:

$$G_{F,n}(x) = Dist_{R_n,F}(x) = P_F(R_n \leq x)$$

– 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

$$\widehat{G}_{F,n}(x) = G_{F_n,n}(x) = Dist^*_{R_n}(x) = Dist_{R_n,F_n}(x) = P_{F_n}(R_n^* \leq x)$$

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

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

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:

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

– 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:

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

• **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!

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

• 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 μ and σ^2 finite.

– In this case:

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

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 intergrable function $g(\cdot)$;
2. the convergence is uniform on $\mathcal{F}$;
3. the mapping $A \to Q(\cdot, A)$ is continuous in A,

then

$\forall x$ and $\forall \epsilon > 0$, $P\left(|Dist_{R_n,F_n}(x) - Q(x, F)| > \epsilon\right) \to 0$ 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
$$Q(x, F) = Q(x) = Expo(1)$$
    - $Dist_{R_n,F_n}$ 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 analyse 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 Γ_p. It is defined by:

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

    where $\mathcal{C}(X, Y)$ 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 Γ_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:

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

  – (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 Γ_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 Γ_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 $X_i \in I\!R^k$ 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:

$$\begin{aligned}
d_2(\hat{K}, K) &= d_2(\sqrt{n}(\bar{X}^* - \bar{X})), \sqrt{n}(\bar{X} - \mu)) \\
&= d_2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i^* - \bar{X}), \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu)\right) \\
&= \frac{1}{\sqrt{n}}\, d_2\left(\sum_{i=1}^{n}(X_i^* - \bar{X}), \sum_{i=1}^{n}(X_i - \mu)\right) \quad \text{by [P3]} \\
&\leq \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}\left[d_2(X_i^* - \bar{X}, X_i - \mu)\right]^2} \quad \text{by [P6]}
\end{aligned}$$

  - Since the sampling is iid the latter is equal to $\frac{\sqrt{n(d_2(X_1^* - \bar{X}, X_1 - \mu))^2}}{\sqrt{n}}$:

$$\begin{aligned}
d_2(X_1^* - \bar{X}, X_1 - \mu) &= d_2(X_1^* - E^*(X_1^*), X_1 - E(X_1)) \\
\text{(by [P4])} \quad &= \sqrt{[d_2(X_1^*, X_1)]^2 - \| E^*(X_1^*) - E(X_1) \|^2} \\
&= \sqrt{[d_2(F_n, F)]^2 - \| \bar{X} - \mu \|^2}
\end{aligned}$$

  - 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:

$$\text{as } n \to \infty,\ d_2(\hat{K}, K) \to 0,\ a.s.$$

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 η, 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 *Attained Significance Level* is

$$p = Prob(T(\mathcal{X}) \geq t \,|\, H_0).$$

– Critical value for a test of size α:

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

$$Prob(T(\mathcal{X}) \geq t_\alpha | H_0) = \alpha$$

– How to compute p-value and/or t_α?

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

– **if** H_0 **is simple**: easy

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

– **if H_0 is composite**: more complicated

– the p-value is not clearly defined because $Prob(T(\mathcal{X}) \geq t_\alpha | F \in H_0)$ 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

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

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

– simulate $\mathcal{X}^{*b}$ under $\widehat{F}_0$, 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 $\widehat{F}_0$?

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 η. 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 η computed under the null $\theta = \theta_0$

  ∗ Under regularity conditions, $\hat{\eta}_0$ converges in probability to a limit η_0 called the pseudo-true value of η

- 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

$$\begin{aligned} H_0 &: f(x) = f_0(x|\zeta) \\ H_1 &: f(x) = f_1(x|\xi) \end{aligned}$$

– 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 likelihhod 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 $t = T_{observed}$.

- **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):

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

– We observe $t = 0.0036$.

– Bootstrap algorithm:

- We generate $B = 1000$ iid sample $\mathcal{X}^*$ of size $n = 10$ from a $Gamma(\hat{\alpha}, \hat{\beta})$
- For each sample $\mathcal{X}^*$ 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

$$f_0(x|\mu) = \frac{e^{-x/\mu}}{\mu}$$
$$f_1(x|\zeta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha}$$

– 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):

$$\begin{aligned} t &= n^{-1}\left\{\ell_1(\hat{\zeta}) - \ell_0(\hat{\mu})\right\} \\ &= (\hat{\alpha}-1)\left(\frac{1}{n}\sum \log(x_i)\right) + \log(\hat{\mu}) + 1 - \log(\Gamma(\hat{\alpha})) - \hat{\alpha} - \hat{\alpha}\log(\hat{\beta}) \end{aligned}$$

– We observe $t = 0.1241$.

– Bootstrap algorithm:

- We generate $B = 1000$ iid sample $\mathcal{X}^*$ of size $n = 10$ from a $Expo(\hat{\mu})$
- For each sample $\mathcal{X}^*$ 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:

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

p-value $= P(\chi^2_{(1)} \geq 0.1241 * 2 * n) = P(\chi^2_{(1)} \geq 2.4818) = 0.1152.$

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 $\widehat{F}_0$ 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:

$$p = Prob(T(\mathcal{X}) \geq t \,|\, H_0, S = s).$$

- 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 difference" in favor of "H_1 : positive effect" if $\hat{\theta}$ is too large.
    - the p-value is

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

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

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

with

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

where $s_p^2 = ((n-1)s_z^2 + (m-1)s_y^2)/(n+m-2)$.

- In **nonparametric** models, for evaluating $P(\hat{\theta} \geq \hat{\theta}_{observed}|H_0)$ 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 distribuion.
- We combine the $N = m + n$ observation in a single sample $\mathcal{X} = (\mathcal{Y}, \mathcal{Z})$
- Under the null, the sufficient statistics is the emprical 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

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

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

- *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

$$\begin{aligned} p - \text{value} &= P(\hat{\theta}(\mathcal{X}) > \hat{\theta}_{observed} \,|\, S = s, H_0) \\ &= P(\hat{\theta}(\mathcal{X}^*) > \hat{\theta}_{observed}) \\ &= \frac{\#\text{ of permutations such that } \hat{\theta}(\mathcal{X}^*) \geq \hat{\theta}_{observed}}{\binom{N}{m}} \end{aligned}$$

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:

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

- **Numerical illustration**

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

$$F_Y = N(10, (1.5)^2), \quad \text{sample size } n = 7$$
$$F_Z = N(12, (1.5)^2), \quad \text{sample size } m = 10$$

- we observe the following statistics

$$\bar{y} = 10.0236 \qquad s_y^2 = 1.7213$$
$$\bar{z} = 11.3601 \qquad s_z^2 = 1.1464$$

- The student statistics ($s_p^2 = 1.3763$ and $\hat{\theta}_{obs} = 1.3365$) takes the value $t_{obs} = 2.3116$.
- The (exact here) p-value is $P(t_{15} \geq t_{obs}) = 0.0177$
- The permutation test on $\hat{\theta} = \bar{Z} - \bar{Y}$ over $B = 1000$ simulation provides

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

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

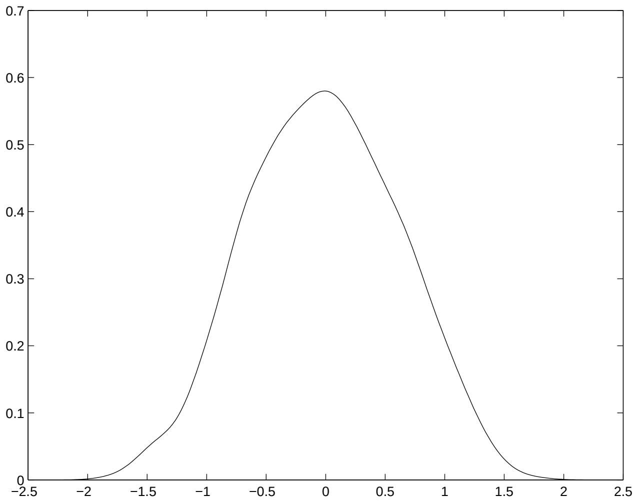

Figure 4.1: *Permutation simulated estimates of the sampling distribution of* $\hat{\theta}(\mathcal{X}^*) = \bar{z}^* - \bar{y}^*$.

– **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 1.0254 5.0447  15.6797)
y=(0.4306  0.1853 0.2734 0.3542  4.7347 0.1250 0.4149)
```

- we have the following statistics

$$\bar{y} = 0.9312 \qquad s_y^2 = 2.8260$$
$$\bar{z} = 6.3535 \qquad s_z^2 = 71.2146$$

- 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

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

- Figure 4.2 illustrates the simulated empirical density of the $B = 1000$ values of $\hat{\theta}(\mathcal{X}^{*(b)})$.
- Here the normal theory with equal variance is not very appropriate.

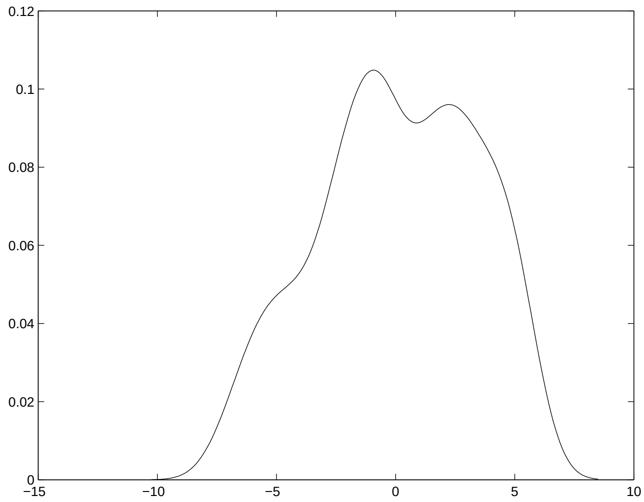

Figure 4.2: *Permutation simulated estimates of the sampling distribution of* $\hat{\theta}(\mathcal{X}^*) = \bar{z}^* - \bar{y}^*$.

• 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:

$$s = \big(y_{(1)}, \ldots, y_{(n)}, z_{(1)}, \ldots, z_{(n)}\big)$$

  – 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

$$\begin{aligned} p - \text{value} &= P(\hat{\rho}(\mathcal{X}) \geq \hat{\rho}_{observed} \,|\, S = s, H_0) \\ &= P(\hat{\rho}(\mathcal{X}^*) \geq \hat{\rho}_{observed}) \\ &= \frac{\#\text{ of permutations such that } \hat{\rho}(\mathcal{X}^*) \geq \hat{\rho}_{observed}}{n!} \end{aligned}$$

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

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

- **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 ploted in Figure 4.3. We obtain the following results from the permutation test ($B = 1000$):

$$\begin{aligned} \hat{\rho} &= 0.0929 \\ \hat{p} &= 0.3337 \end{aligned}$$

The right panel of Figure 4.3 provides the sampling distribution of $\hat{\rho}(\mathcal{X}^*)$ under H_0.

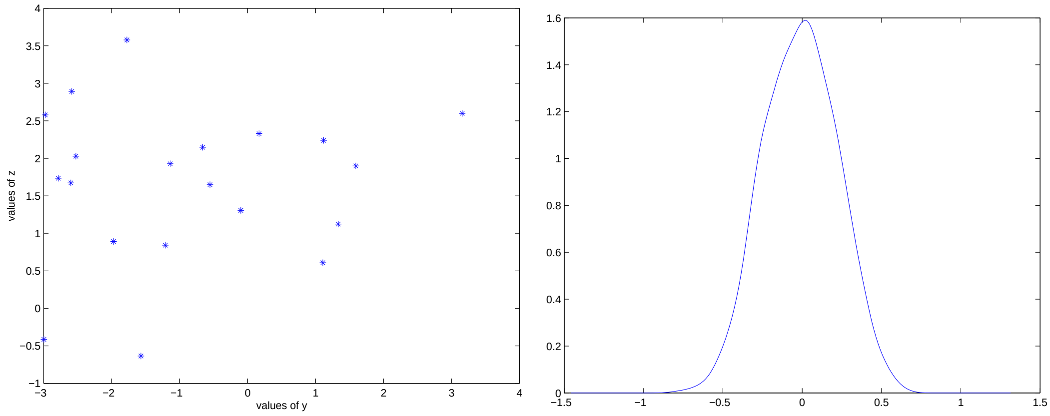

Figure 4.3: *Two independent normal variates. Left panel: a simulated sample ($n = 20$). Right panel: simulated estimates of the sampling distribution of $\hat{\rho}(\mathcal{X}^*)$ under H_0. Here,* $\hat{\rho} = 0.0929$

– **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 ploted 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 $\hat{\rho}(\mathcal{X}^*)$ under H_0.

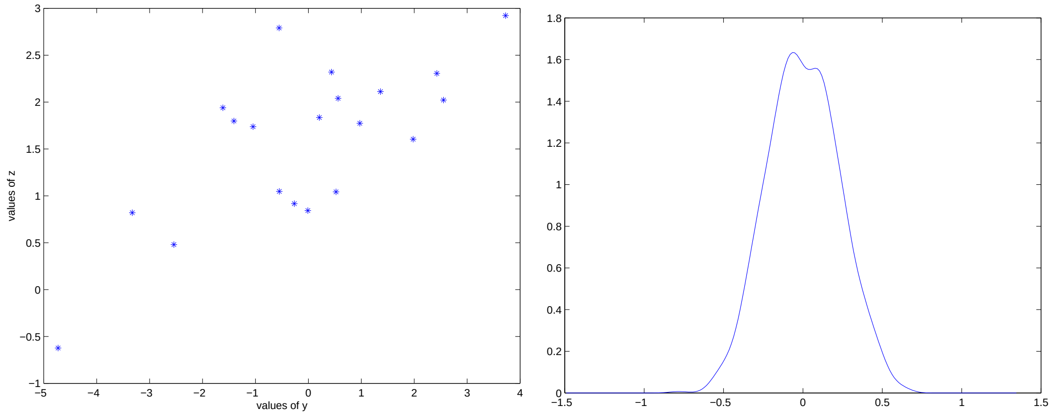

Figure 4.4: *A random sample from bivariate normal with* $\rho = 0.50$*. Left panel: a simulated sample (*$n = 20$*). Right panel: simulated estimates of the sampling distribution of* $\hat{\rho}(\mathcal{X}^*)$ *under* H_0*. Here,* $\hat{\rho} = 0.7205$

– **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 ploted 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 $\hat{\rho}(\mathcal{X}^*)$ under H_0.

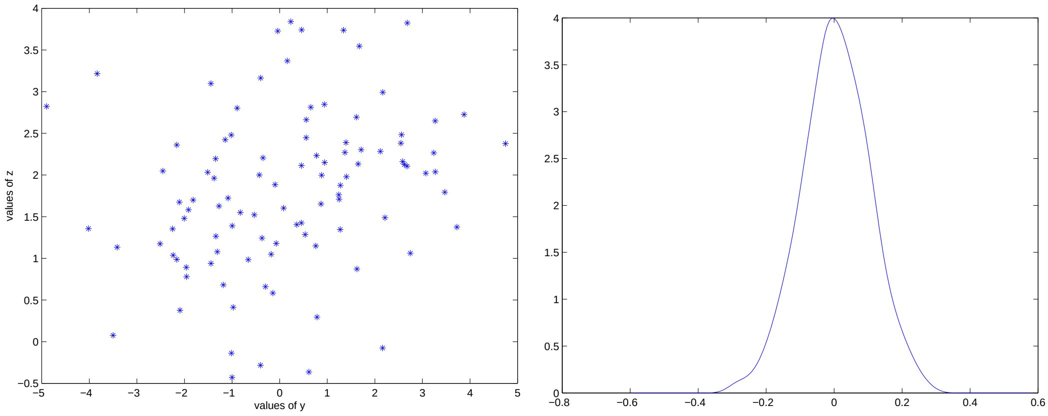

Figure 4.5: *A random sample from bivariate normal with* $\rho = 0.25$*. Left panel: a simulated sample (*$n = 100$*). Right panel: simulated estimates of the sampling distribution of* $\hat{\rho}(\mathcal{X}^*)$ *under* H_0*. Here,* $\hat{\rho} = 0.2506$*.*

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 $\widehat{F}_0$ which satisfies H_0: it cannot be simply F_n.

- Depends on the problem
- Many candidates here for $\widehat{F}_0$, 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:

$$\begin{aligned} \hat{p} &= P(T(\mathcal{X}^*) \geq T_{observed} \,|\, \widehat{F}_0) \\ &\approx \frac{1 + \#\{T(\mathcal{X}^{(*b)}) \geq T_{observed}\}}{B+1} \end{aligned}$$

• **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):

$$t = \frac{\bar{z} - \bar{y}}{\sqrt{s_Z^2/n + s_Y^2/m}},$$

– F_N, the empirical cdf of $\mathcal{X}$, is a natural estimator of the common cdf of Y and Z.

– **Numerical illustration**:

- 2 independent samples from unknown populations (see above)

```
z=(2.5884  1.4106 8.6561 1.4820 26.1792 0.7062 0.7625 1.0254 5.0447  15.6797)
y=(0.4306  0.1853 0.2734 0.3542  4.7347 0.1250 0.4149)
```

- we had the following statistics

$$\begin{aligned} \bar{y} &= 0.9312 \qquad & s_y^2 &= 2.8260 \\ \bar{z} &= 6.3535 \qquad & s_z^2 &= 71.2146 \\ \hat{\theta}_{obs} &= 5.4223 \end{aligned}$$

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

– The boostrap test, with $B = 1000$ replications, provides

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

– Figure 4.6 illustrates the simulated empirical density of the $B = 1000$ values of $\hat{\theta}(\mathcal{X}^{*(b)})$.

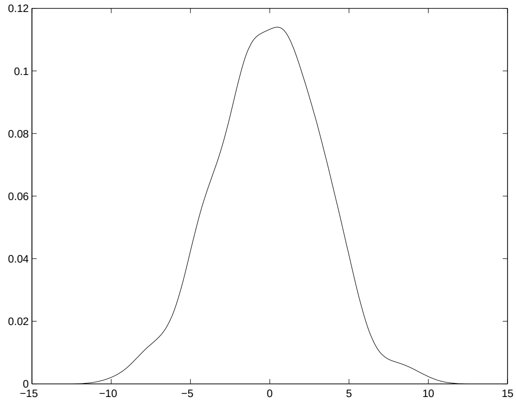

Figure 4.6: *Bootstrap estimates of the sampling distribution of* $\hat{\theta}(\mathcal{X}^*) = \bar{z}^* - \bar{y}^*$.

– 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 squarred 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

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

  – How to select $\widehat{F}_0$?

- It cannot be F_n: it does not obey H_0.
- We **shift** the empirical cdf so that it has mean μ_0

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

  $\widehat{F}_0$ is the empirical cdf of the $\tilde{x}$'s.
- The bootstrap test statistics is then

$$T(\tilde{\mathcal{X}}^*) = \frac{\bar{\tilde{X}}^* - \mu_0}{\tilde{S}^*/\sqrt{n}}$$

  where $\bar{\tilde{X}}^*$ and $\tilde{S}^*$ are the sample mean and std deviation of the bootstrap samples $\tilde{\mathcal{X}}^*$ obtained by drawing with replacement in the shifted values $\tilde{\mathcal{X}}$.

– The p-value is then obtained by

$$\hat{p} = P(T(\tilde{\mathcal{X}}^*) \geq T_{observed} \,|\, \widehat{F}_0)$$

– Note that in this case, the bootstrap statistics can also be calculated without explicitely shifting:

$$T(\tilde{\mathcal{X}}^*) \equiv \frac{\bar{X}^* - \bar{X}}{S^*/\sqrt{n}}$$

where $\bar{X}^*$ and S^* are the sample mean and std deviation of the usual bootstrap samples $\mathcal{X}^*$.

- **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 $P(\hat{\rho}(\mathcal{X}) \geq \hat{\rho}_{obs} | H_0)$ (one-sided case).
  - The bootstrap estimate of the p-value is obtained as

$$\hat{p} = P(\hat{\rho}(\mathcal{X}^*) \geq \hat{\rho}_{obs} | H_0)$$

    where $\mathcal{X}^* = \{\mathcal{Y}^*, \mathcal{Z}^*\}$ with $\mathcal{Y}^*$ and $\mathcal{Z}^*$ being **independent** bootstrap samples from $\mathcal{Y}$ and $\mathcal{Z}$ respectively.
  - In practice we simulate B bivariate random samples $\mathcal{X}^{*(b)}$ and calculate $\hat{\rho}(\mathcal{X}^{*(b)})$, $b = 1, \ldots, B$ and approximate the p-value by:

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

  - 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 ($\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = 2$ and $\sigma_2 = 1$). The obtained sample is ploted 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 $\hat{\rho}(\mathcal{X}^*)$ under H_0.

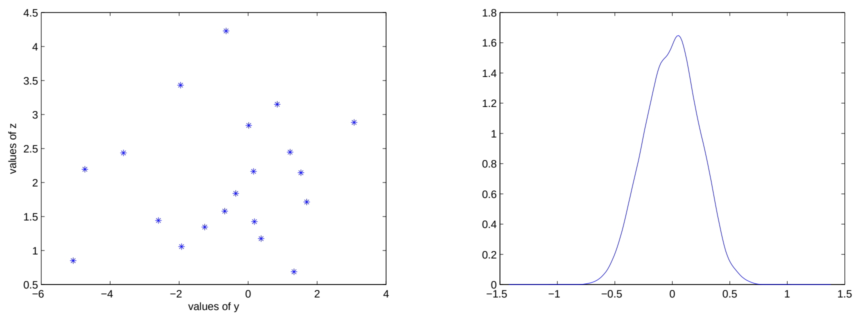

Figure 4.7: *Two independent normal variates. Left panel: a simulated sample ($n = 20$). Right panel: simulated estimates of the sampling distribution of $\hat{\rho}(\mathcal{X}^*)$ under H_0. Here,* $\hat{\rho} = 0.1584$

– **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 ploted 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 $\hat{\rho}(\mathcal{X}^*)$ under H_0.

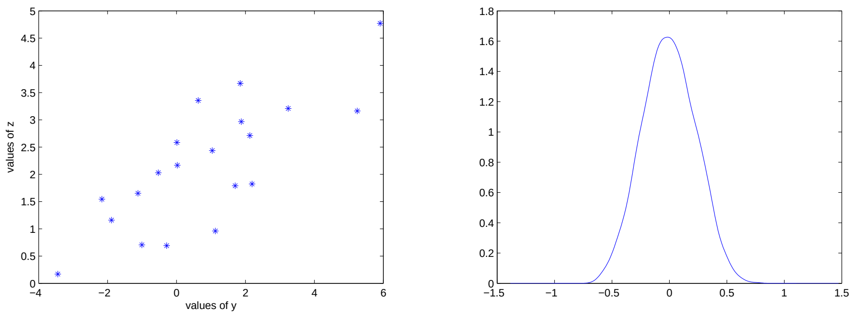

Figure 4.8: *A random sample from bivariate normal with $\rho = 0.50$. Left panel: a simulated sample ($n = 20$). Right panel: simulated estimates of the sampling distribution of $\hat{\rho}(\mathcal{X}^*)$ under H_0. Here, $\hat{\rho} = 0.7823$*

– **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 ploted 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.

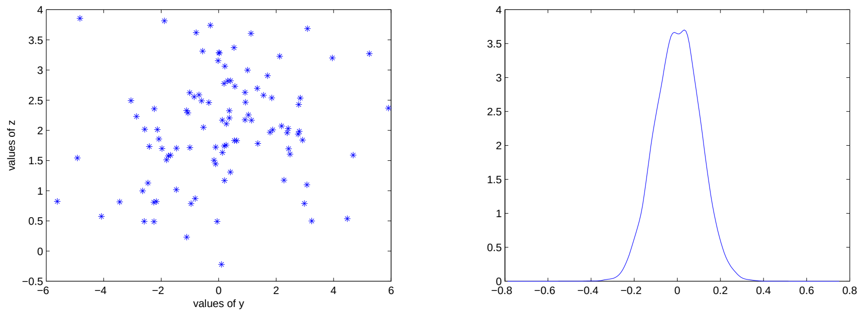

Figure 4.9: *A random sample from bivariate normal with $\rho = 0.25$. Left panel: a simulated sample ($n = 100$). Right panel: simulated estimates of the sampling distribution of $\hat{\rho}(\mathcal{X}^*)$ under H_0. Here, $\hat{\rho} = 0.1820$.*

– **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.

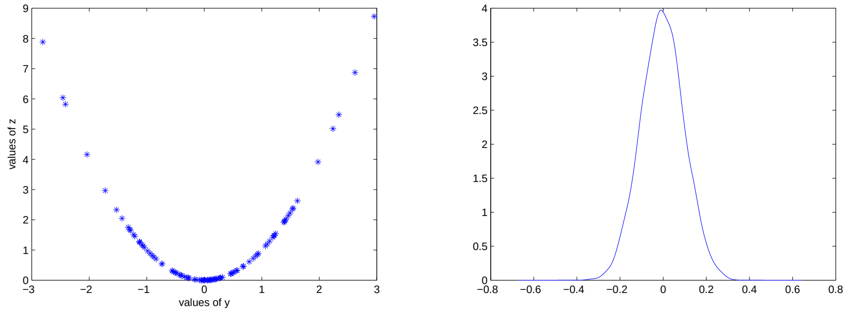

Figure 4.10: *Left panel: a simulated sample ($n = 100$). Right panel: simulated estimates of the sampling distribution of $\hat{\rho}(\mathcal{X}^*)$ under H_0. Here, $\hat{\rho} = 0.0953$.*

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 ith 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 ϵ_i are i.i.d. with common cdf F with mean 0 and constant variance σ^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:

$$\sqrt{n}(\hat{\beta} - \beta) \to N_p(0, \sigma^2 V^{-1}).$$

– 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 β_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:

$$\begin{aligned}\hat{\epsilon} &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= M_X\, y\end{aligned}$$

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 ϵ:

$$E_F(\hat{\epsilon}) = 0$$
$$Var_F(\hat{\epsilon}) = M_X(\sigma^2 I_n)M_X = \sigma^2 M_X = \sigma^2(I_n - H_X)$$

– 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,

$$y_i = x_i\beta + \epsilon_i, \quad \text{where } \epsilon_i \sim F(0, \sigma^2).$$

This provides the original sample: $\mathcal{X} = \{(x_i, y_i) \mid i = 1, \dots, n\}$.

• In the ***bootstrap world*** the data generating process is

$$\widehat{\mathcal{P}} = (\hat{\beta}, F_n)$$

where F_n is either the empirical cdf of the centered residuals (nonparametric 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 residulas, $\Rightarrow$ center the residuals:

$$\tilde{\epsilon}_i^{(1)} = \hat{\epsilon}_i - \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i$$

– We define here F_n as the empirical cdf of $\tilde{\epsilon}_i^{(1)}$, $i = 1, \dots, n$.

– Draw independently from F_n to produce the bootstrap residuals ϵ_i^*, $i = 1, \dots, n$.

– Generate with $\widehat{\mathcal{P}}$ the bootstrap values:

$$y_i^* = x_i\hat{\beta} + \epsilon_i^*, \quad i = 1, \ldots, n$$

– The pseudo-sample $\mathcal{X}^* = \{(x_i, y_i^*) \,|\, i = 1, \ldots, n\}$ provides the bootstrap value of $\hat{\beta}$:

$$\hat{\beta}^* = (X'X)^{-1}X'y^*$$

– The bootstrap analog of the root for β_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

$$\hat{\sigma}^{*2} = \frac{1}{n-p}(y^* - X\hat{\beta}^*)'(y^* - X\hat{\beta}^*).$$

– In particular, the bootstrap-t confidence interval for β_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 $Dist_{U_j}^*(x)$.

• It can be proven that the bootstrap-t achieve 3rd order accuracy for the slopes parameters ($O_p(n^{-3/2})$).

- Additional refinement:
  - Since $Var_F(\hat{\epsilon}_i) = (1 - h_i)\sigma^2 \neq \sigma^2$, we may also correct for the second moment and then center the obtained residuals.
  - Define the corrected residuals:

$$\tilde{\epsilon}_i^{(2)} = \frac{\hat{\epsilon}_i}{\sqrt{1 - h_i}} - \sum_{j=1}^{n} \frac{\hat{\epsilon}_j}{\sqrt{1 - h_j}}$$

  - Now we define F_n as the empirical cdf of $\tilde{\epsilon}_i^{(2)}, \ i = 1, \ldots, n$.

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:

$$\epsilon_i \sim Gamma(a, c) - \mu$$

    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).

sample size	$n = 5$		$n = 10$		$n = 20$		$n = 100$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	5.505	0.855	4.127	0.918	2.917	0.931	1.350	0.950
Perc. Boot.	4.210	0.763	3.695	0.881	2.763	0.913	1.337	0.948
Bootstrap-t	8.620	0.941	4.824	0.953	3.096	0.943	1.363	0.952

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.

sample size	$n = 5$		$n = 10$		$n = 20$		$n = 100$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	5.350	0.854	4.104	0.909	2.964	0.940	1.346	0.943
Perc. Boot.	5.232	0.849	4.106	0.905	2.967	0.941	1.344	0.944
Bootstrap-t	8.424	0.954	4.794	0.946	3.158	0.960	1.358	0.944

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.

Sample size	Min	Max
$n = 5$	0.63	0.90
$n = 10$	0.81	0.95
$n = 20$	0.90	0.97
$n = 100$	0.98	0.995

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:

$$\forall i, \quad E(x_i'\epsilon_i) = 0$$

– We need also an additional regularity assumptions on X to get the asymptotics:

$$\forall i, \quad E(x_i'x_i) = V > 0$$

– 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

$$\hat{\beta}^* = (X^{*\prime} X^*)^{-1} X^{*\prime} y^*,$$

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

$$\hat{\sigma}^{*2} = \frac{1}{n-p}(y^* - X^*\hat{\beta}^*)'(y^* - X^*\hat{\beta}^*)$$

  - The bootstrap analog of the root for β_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 $(X^{*\prime} X^*)^{-1}$.

– The bootstrap-t confidence interval for β_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 $X^{*\prime}X^{*}$ 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).

sample size		$n = 10$		$n = 20$		$n = 100$		$n = 500$
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	4.441	0.923	3.085	0.934	1.358	0.942	0.606	0.956
Perc. Boot.	3.976	0.889	2.933	0.918	1.343	0.940	0.604	0.952
Bootstrap-t	5.188	0.952	3.281	0.951	1.369	0.944	0.606	0.953

Table 5.4: *Performances of confidence intervals for $\beta_2 = 1$, Bootstrapping on the residual. Same senario as in Table 5.1 but here, $X \sim U(0,1)$, over $M = 1000$ Monte-Carlo experiments. For the bootstrap, $B = 2000$.*

sample size	$n = 10$		$n = 20$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	4.425	0.917	3.053	0.936	1.360	0.947	0.607	0.961
Perc. Boot.	4.725	0.911	3.011	0.922	1.346	0.941	0.605	0.956
Bootstrap-t	6.273	0.951	3.356	0.926	1.369	0.939	0.606	0.956

Table 5.5: *Performances of confidence intervals for* $\beta_2 = 1$*, Bootstrapping on the pairs. Same senario 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:

$$y = 1 + 2x + \epsilon \text{ with } \epsilon \sim N(0, \sigma^2(x)),$$

  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 $n \geq 100$).

sample size	$n = 10$		$n = 20$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	5.038	0.868	3.532	0.913	1.567	0.921	0.701	0.919
Perc. Boot.	4.507	0.827	3.359	0.887	1.550	0.918	0.698	0.917
Bootstrap-t	5.858	0.919	3.761	0.925	1.581	0.919	0.700	0.920

Table 5.6: *Performances of confidence intervals for* $\beta_2 = 2$*, Bootstrapping on the residual, over* $M = 1000$ *Monte-Carlo experiments. For the bootstrap,* $B = 2000$*.*

sample size	$n = 10$		$n = 20$		$n = 100$		$n = 500$	
method	length	coverage	length	coverage	length	coverage	length	coverage
As. Normal	5.038	0.859	3.568	0.911	1.564	0.936	0.701	0.929
Perc. Boot.	5.486	0.882	3.688	0.915	1.677	0.952	0.764	0.957
Bootstrap-t	8.609	0.913	4.443	0.917	1.725	0.951	0.767	0.956

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 ϵ_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 $\widehat{\Omega}$ (heteroskedastic consistent covariance matrix estimator) of Ω. Basic estimator of $Var(\hat{\beta}_{OLS})$ is given by

$$Var(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\widehat{\Omega}X(X'X)^{-1},$$

where typically $\widehat{\Omega} = \text{diag}(\hat{u}_1^2, \ldots, \hat{u}_n^2)$ 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 $u_i^* = \hat{u}_i\varepsilon_i$ where ε_i are mutually independent drawings completely independent of the original data and such that $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = 1$ and $E(\varepsilon_i^3) = 1$. Mammen (1993) suggests the use of a two-point distribution for ε_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:

$$W_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{cov_{jj}}}, \quad \text{and} \quad W_j^* = \frac{\hat{\beta}_j^* - \hat{\beta}_j}{\sqrt{cov_{jj}^*}},$$

where cov_{jj} is the (j,j)th element of the matrix $(X'X)^{-1}X'\widehat{\Omega}X(X'X)^{-1}$ and cov_{jj}^* is the (j,j)th element of the matrix $(X'X)^{-1}X'\widehat{\Omega}^*X(X'X)^{-1}$.

- **More Monte-Carlo Examples with Wild Bootstrap**

<u>Homoskedastic case</u>: simulated data according the model:

– The regression model is

$$y_i = 2 + x_i + \epsilon_i$$

where $\epsilon_i \sim \Gamma(a,c) - \mu$, with $\mu = ac$, $a = 2, c = 1/ < sqrt2$, so $\sigma_\epsilon = 1$.

– X's are random variables $X \sim U(1,20)$.

– We do not analyze the case $n = 5$ for boostrapping on the pairs (drawing 5 rows among 5 original rows produces too many singular $X^{*\prime}X^*$).

– Table 5.8 show the results

sample size		$n = 5$		$n = 10$		$n = 20$		$n = 100$
method	length	coverage	length	coverage	length	coverage	length	coverage
OLS As. N.	0.372	0.838	0.235	0.915	0.163	0.934	0.072	0.946
Perc. (Res)	0.284	0.753	0.210	0.885	0.155	0.922	0.071	0.944
Boot-t (Res)	0.627	0.937	0.275	0.954	0.174	0.945	0.073	0.950
Perc. (Pairs)	n.a.	n.a.	0.255	0.903	0.160	0.919	0.071	0.943
Boot-t (Pairs)	n.a.	n.a.	0.333	0.946	0.178	0.943	0.073	0.948
HCCME As. N.	0.229	0.684	0.190	0.849	0.146	0.907	0.071	0.944
Perc. (Wild)	0.225	0.671	0.187	0.834	0.145	0.895	0.070	0.938
Boot-t (Wild)	0.533	0.828	0.251	0.887	0.160	0.921	0.072	0.941

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).

sample size		$n = 5$		$n = 10$		$n = 20$		$n = 100$
method	length	coverage	length	coverage	length	coverage	length	coverage
OLS As. N.	6.742	0.734	4.623	0.820	3.282	0.806	1.560	0.852
Perc. (Res)	5.127	0.626	4.177	0.772	3.145	0.782	1.553	0.844
Boot-t (Res)	11.139	0.854	5.362	0.869	3.480	0.832	1.575	0.857
Perc. (Pairs)	n.a.	n.a.	4.964	0.875	3.854	0.897	1.994	0.945
Boot-t (Pairs)	n.a.	n.a.	10.976	0.929	5.477	0.924	2.109	0.939
HCCME As. N.	4.725	0.590	4.668	0.811	3.831	0.866	2.038	0.932
Perc. (Wild)	4.444	0.569	4.504	0.781	3.743	0.833	2.017	0.919
Boot-t (Wild)	11.838	0.792	6.757	0.895	4.350	0.881	2.046	0.920

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

$$T(\mathcal{X}) = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}$$

where c_{jj} is the (j, j)th element of the matrix $(X'X)^{-1}$.

– The p-value is given by:

$$p = P(T(\mathcal{X}) \geq T_{observed} \,|\, H_0)$$

– We have to estimate p by the bootstrap algorithm. How to generate the samples under H_0?

- We estimate β 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 jth column deleted.

- We define the residuals under H_0 as

$$\hat{\epsilon}^{(0)} = y - \tilde{X}\hat{\beta}^{(0)}$$

- We draw with replacement n values from $\hat{\epsilon}_i^{(0)},\ i = 1, \ldots, n$ to produce the bootstrap residuals $\hat{\epsilon}_i^{(0),*},\ i = 1, \ldots, n$.

– The bootstrap sample under H_0 is $\mathcal{X}^{(0),*} = \{(x_i, y_i^*)|i = 1, \ldots, n\}$ where

$$y_i^* = \tilde{x}_i\,\hat{\beta}^{(0)} + \hat{\epsilon}_i^{(0),*}.$$

– 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 jth 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 ($n = 8$):

```
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 data are presented in Figure 5.1 with the least squares fit.

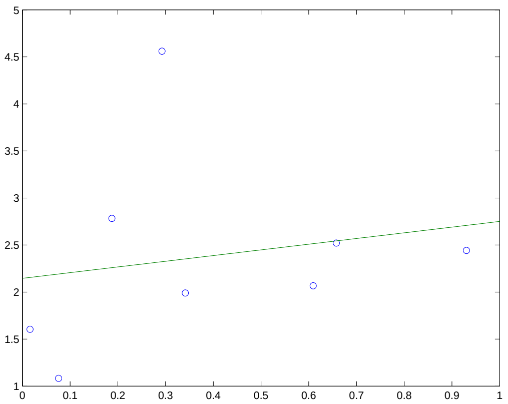

Figure 5.1: *Scatterplot of the data and OLS fit.*

  - The OLS fit gives $\hat{\beta} = (2.1458 \quad 0.6048)'$, 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 β_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:
$$y_f = x_f\beta + \epsilon_f, \quad \text{where } \epsilon_f \sim F(0, \sigma^2).$$
  - The prediction is given by
$$\hat{y}_f = x_f\hat{\beta}$$
where $\hat{\beta}$ is the estimator of β 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:
$$\begin{aligned} E_F(e_f) &= 0 \\ Var_F(e_f) &= \sigma^2\left(1 + x_f(X'X)^{-1}x_f'\right) = \sigma^2(1 + h_f) \end{aligned}$$
  - 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 $\widehat{\mathcal{P}}$ is characterized by:

$$y = x\hat{\beta} + \epsilon^*, \quad \text{where } \epsilon^* \sim F_n$$

– y_f^* is generated by $\widehat{\mathcal{P}}$ 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 $\mathcal{X}^*$ of size n generated from $\widehat{\mathcal{P}}$.

– The error of prediction is

$$e_f^* = y_f^* - \hat{y}_f^* = x_f(\hat{\beta} - \hat{\beta}^*) + \epsilon_f^*$$

– 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 $\mathcal{X}^*$.

– 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 $Dist^*_R$.

- 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 $\mathcal{X}^*$ 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 ϵ^*_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^*: $Dist^*_R$, 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 ($B = 5000$):

$$\begin{array}{ll} \text{Asymptotic Normal:} & [0.1501,\ 4.7463] \\ \text{Nonparametric Bootstrap:} & [-3.3438,\ 4.4533] \end{array}$$

  - *Remark: This sample was obtained by simulation $n = 8$ values of $X \sim U(0, 1)$ and $\epsilon \sim Shifted - Gamma$ 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 $\sigma^2(F) = n\, AVar_F(T(\mathcal{X}))$ 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 consitent 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 $\hat{\sigma}^{*2}$

$$\hat{\sigma}^{*2} = n\, Var_{F_n}(T(\mathcal{X}^*))$$

  - 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 $Dist_W(x)$ and $\hat{\sigma}^*$ is the first level bootstrap estimate of σ.

    - Since $Dist_W(x)$ 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 $(\hat{\sigma}^{**})^2 = n\, Var_{F_n^*}(T(\mathcal{X}^{**}))$ is a bootstrap estimator in the *bootstrap world* of $\hat{\sigma}^{*2} = n\, Var_{F_n}(T(\mathcal{X}^*))$.
      * 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 σ.

– 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:

$$\hat{\sigma}^{**(b_1)} = \sqrt{n}\left(Var^{**}(T^{*(b_1)})\right)^{\frac{1}{2}}$$

[3] Compute the bootstrap value of the root:

$$W^{*(b_1)} = \sqrt{n}\left(\frac{T(\mathcal{X}^{*(b_1)}) - \theta(F_n)}{\hat{\sigma}^{**(b_1)}}\right)$$

[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 σ:

$$\hat{\sigma}^* = \sqrt{n}\left(\frac{1}{B_1}\sum_{b_1=1}^{B_1} T^2(\mathcal{X}^{*(b_1)}) - \left(\frac{1}{B_1}\sum_{b_1=1}^{B_1} T(\mathcal{X}^{*(b_1)})\right)^2\right)^{\frac{1}{2}}$$

• The 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 $Dist_W^*(x)$.

• Remark : Computer intensive method : $B_1 * B_2$ Monte-Carlo loops where both B_1 should be, say ≥ 2000 and B_2 should be, say ≥ 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 σ.

– Suppose instead, we use the double bootstrap to estimate σ. 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).

method	lower limit	upper limit
Usual bootstrap-t	25.1980	300.7924
Double bootstrap-t	24.0624	288.3107

Table 6.1: *Confidence intervals, duration data with $n = 10$. Here $B1 = 2000$ and $B2 = 200$.*

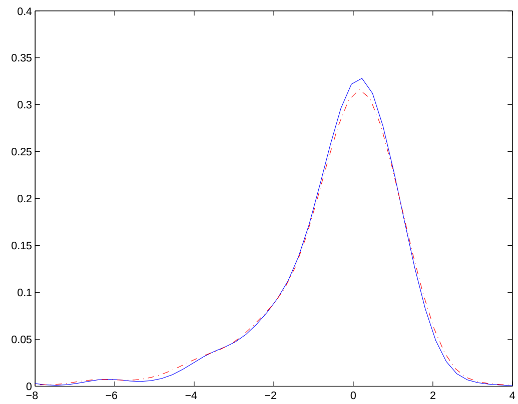

Figure 6.1: *Empirical bootstrap densities of the two roots: usual studentized (solid) and bootstrap studentized (dash-dotted).*

- Part of the Matlab code for the double bootstrap:

```
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 $\theta(F)$:

$$\mathcal{I}_\alpha = [T(\mathcal{X}) - w^*(1 - \frac{\alpha}{2}), T(\mathcal{X}) - w^*(\frac{\alpha}{2})]$$

  - 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

$$\mathcal{X}^{**} = (X_1^{**}, \dots, X_n^{**})$$

where X_i^{**} is drawn with replacement from $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$ for $i = 1, \dots, n$.

– Here we have $X_i^{**} \sim F_n^*$ where F_n^* is the empirical cdf of $\mathcal{X}^* = (X_1^*, \dots, X_n^*)$.

– 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!**):

$$\mathcal{I}_\alpha^* = [T(\mathcal{X}^*) - w^{**}(1-\frac{\alpha}{2}), T(\mathcal{X}^*) - w^{**}(\frac{\alpha}{2})]$$

– Here $w^{**}(a)$ is the $a-$quantile of $Dist_W^{**}(x)$, the empirical cdf of W^{**} obtained through the B_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:

$$\mathcal{I}_\alpha^* = \mathcal{I}_\alpha^{*(b_1)}.$$

– Since $\theta(F_n)$ is known, we can estimate $\pi(\alpha)$ by the observed proportion of time $\theta(F_n) \in \mathcal{I}_\alpha^*$:

$$\hat{\pi}(\alpha) = P_{F_n}(\theta(F_n) \in \mathcal{I}_\alpha^*)$$

– This is obtained through the following:

$$\hat{\pi}(\alpha) \approx \frac{1}{B_1}\sum_{b_1=1}^{B_1} I(\theta(F_n) \in \mathcal{I}_\alpha^{*(b_1)})$$

– This is the proportion of times that $\mathcal{I}_\alpha^*$ covers $\theta(F_n)$ in repeating both levels of the Bootstrap many times.

- **Calibration**

  – Solve in α

  $$\hat{\pi}(\alpha) = 1 - \alpha_0$$

  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 $\mathcal{I}_{\hat{\alpha}}$:

  $$\mathcal{I}_{\hat{\alpha}} = [T(\mathcal{X}) - w^*(1 - \frac{\hat{\alpha}}{2}), T(\mathcal{X}) - w^*(\frac{\hat{\alpha}}{2})]$$

  where $w^*(a)$ is, as above, the $a-$quantile of $Dist_W^*(x)$

  – The true coverage probability is $\pi(\hat{\alpha})$

  $$\pi(\hat{\alpha}) = P_F(\theta(F) \in \mathcal{I}_{\hat{\alpha}}) \approx 1 - \alpha_0$$

  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 $\mathcal{X}^{*(b_1)}$ from $\mathcal{X}$: this provides $T^{*(b_1)} = T(\mathcal{X}^{*(b_1)})$. Denote $F_n^{*(b_1)}$ the empirical cdf of the sample $\mathcal{X}^{*(b_1)}$.

  [2] Estimate a confidence interval $\mathcal{I}_\alpha^{*(b_1)}$ for $\theta(F_n)$ from this estimator $T(\mathcal{X}^{*(b_1)}) \Rightarrow$ *second level bootstrap*:

-[2.1] Draw a sample $\mathcal{X}^{**(b_2)}$ from $\mathcal{X}^{*(b_1)}$, compute $T(\mathcal{X}^{**(b_2)})$, so:

$$W^{**(b1)} = T(\mathcal{X}^{**(b_2)}) - \theta(F_n^{*(b_1)})$$

-[2.2] Redo the step [2.1] for $b_2 = 1, \ldots, B_2$ and form the empirical cdf of $W^{**(b1)}$ 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:

$$\mathcal{I}_\alpha^{*(b_1)} = \left[T(\mathcal{X}^{*(b_1)}) - w^{**(b_1)}(1-\frac{\alpha}{2}), T(\mathcal{X}^{*(b_1)}) - w^{**(b_1))}(\frac{\alpha}{2})\right].$$

[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 $Dist_W^*(x)$ **and** B_1 values of $\mathcal{I}_\alpha^{*(b_1)}$.

[6] Compute

$$\hat{\pi}(\alpha) = \frac{1}{B_1}\sum_{b_1=1}^{B_1} I(\theta(F_n) \in \mathcal{I}_\alpha^{*(b_1)})$$

[7] Solve in α 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 θ is

$$\mathcal{I}_{\hat{\alpha}} = [T(\mathcal{X}) - w^*(1-\frac{\hat{\alpha}}{2}), T(\mathcal{X}) - w^*(\frac{\hat{\alpha}}{2})]$$

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 α_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 : $B_1 * B_2$ Monte-Carlo loops where B_1 and B_2 should be large (say, both ≥ 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 μ: $\mu \in [9.4834,\ 10.9626]$. 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 α_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 μ 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.

$1-\alpha$	$\hat{\pi}(\alpha)$	confidence intervals
0.8000	0.7650	[9.7577, 10.7239]
0.8500	0.8190	[9.6943, 10.7874]
0.9000	0.8770	[9.6163, 10.8631]
0.9100	0.8850	[9.5757, 10.8766]
0.9200	0.8990	[9.5552, 10.8915]
0.9300	0.9110	[9.5369, 10.9205]
0.9400	0.9260	[9.5076, 10.9369]
0.9500	0.9360	[9.4834, 10.9626]
0.9600	0.9430	[9.4421, 10.9878]
0.9700	0.9570	[9.4091, 11.0479]
0.9800	0.9660	[9.2715, 11.1204]
0.9900	0.9810	[9.2378, 11.1828]

Table 6.2: *Nominal levels and estimated coverages. Here* $B1 = 1000$ *and* $B2 = 1000$.

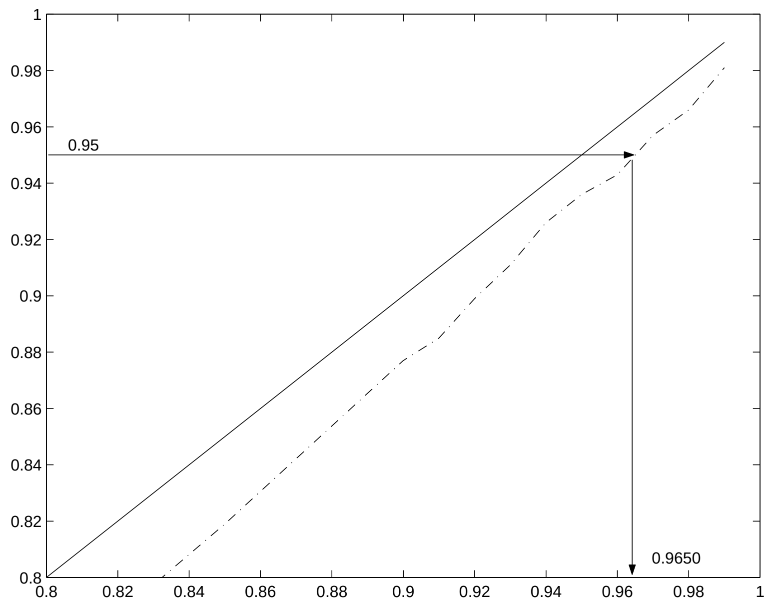

Figure 6.2: *Horizontal scale: nominal levels* $1-\alpha$ *and vertical scale estimated coverage* $\hat{\pi}(\alpha)$. *Here, for* $\alpha_0 = 0.05$, *the calibrated* α *is* $\hat{\alpha} = 1 - 0.9650 = 0.0350$.

• 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 θ: $T_n = \theta(F_n)$).

- 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" $T_{n,i}$

$$T_{n,i} = n\,T_n - (n-1)\,T_{n-1}^{(i)}$$

Note that, in the particular case $T_n(\mathcal{X}) = \bar{X}$, we have

$$\begin{aligned} T_{n,i} &= X_i \\ \bar{X} &= \frac{1}{n}\sum_{i=1}^{n} T_{n,i}. \end{aligned}$$

– The **Jacknife estimators**:

- Jacknife estimator of the mean of T_n:

$$\begin{aligned} \widehat{E}_{\text{jack}}(T_n) &= \widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{n,i} \\ &= nT_n - \frac{n-1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)} \end{aligned}$$

This is the average of the n pseudo-values $T_{n,i}$

- Jacknife estimator of the bias of T_n:

$$\begin{aligned} \widehat{Bias}_{\text{jack}}(T_n) &= T_n - \widehat{E}_{\text{jack}}(T_n) \\ &= (n-1)\left[\frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)} - T_n\right]. \end{aligned}$$

- Jacknife estimator of the variance of T_n:

$$\begin{aligned} \widehat{Var}_{\text{jack}}(T_n) &= \frac{1}{n}\,\widehat{S}_n^2 \\ &= \frac{1}{n}\,\left\{\frac{1}{n-1}\sum_{i=1}^{n}(T_{n,i} - \widehat{T}_n)^2\right\} \end{aligned}$$

$\widehat{S}_n^2$ 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 $\widehat{T}_n = nT_n - \frac{n-1}{n}\sum_{i=1}^n T_{n-1}^{(i)}$, we have:

$$\begin{aligned} E(\widehat{T}_n) &= n\,E(T_n) - (n-1)\,E(T_{n-1}^{(i)}) \\ &= n\,(\theta + \frac{a(\theta)}{n} + \frac{b(\theta)}{n^2} + \ldots) \\ &\qquad -(n-1)\,(\theta + \frac{a(\theta)}{n-1} + \frac{b(\theta)}{(n-1)^2} + \ldots) \\ &= \theta + \frac{b(\theta)}{n} - \frac{b(\theta)}{n-1} + \ldots \\ &= \theta - \frac{b(\theta)}{n(n-1)} + \ldots \end{aligned}$$

  – $\widehat{T}_n$ should be a more accurate estimator of θ

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

$$\widehat{T}_n = \bar{X}_n = T_n$$
$$\widehat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = S_n^2$$

We obtain

$$\begin{aligned}\widehat{Bias}_{\text{jack}}(\bar{X}_n) &= 0\\ \widehat{Var}_{\text{jack}}(\bar{X}_n) &= \frac{S_n^2}{n}\end{aligned}$$

- **The variance**
  - Here $\theta(F) = Var(X)$ 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:

$$\widehat{T}_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = S_n^2$$

    - Note that $E(T_n) = \frac{n-1}{n}\theta$ and $E(\widehat{T}_n) = \theta$.

- **Failure of the jacknife**

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:

$$T_{n,i} = \begin{cases} nT_n - (n-1)X_{(m+1)} & \text{if } i = (1), \ldots, (m) \\ nT_n - (n-1)X_{(m)} & \text{if } i = (m+1), \ldots, (n). \end{cases}$$

- The estimation of the mean of T_n is:

$$\begin{aligned} \widehat{T}_n &= nT_n - \frac{m(n-1)}{n} X_{(m+1)} - \frac{m(n-1)}{n} X_{(m)} \\ &= nT_n - (n-1)T_n = T_n \end{aligned}$$

– Jacknife estimation of the bias:

$$\widehat{Bias}_{\text{jack}}(T_n) = T_n - \widehat{T}_n = 0$$

– But the jacknife estimation of the variance does not work!

- It can be shown that

$$\begin{aligned} \widehat{Var}_{\text{jack}}(T_n) &= \frac{1}{n}\, \widehat{S}_n^2 \\ &= \frac{n-1}{4}(X_{(m+1)} - X_{(m)})^2 \end{aligned}$$

- Whereas (for symetric densities):

$$AVar(T_n) = \frac{1}{4\, n\, f_X^2(\theta)}$$

- The two quantities are different even when $n \to \infty$:

$$\frac{1}{n}\, \widehat{S}_n^2 - \frac{1}{4\, n\, f_X^2(\theta)} \not\to 0 \quad \text{as } n \to \infty$$

– **The boundary of a support:**

- Here $\theta = \inf\{x|F(x) = 1\}$ and $T_n = X_{(n)}$.
- One can show that

$$\widehat{T}_n = X_{(n)} + \frac{n-1}{n}(X_{(n)} - X_{(n-1)})$$

- One can prove that:

$$E(T_n) = \theta + O(n^{-1})$$
$$E(\widehat{T}_n) = \theta + O(n^{-2})$$

– But here again, the jacknife estimation of the variance does not work!

7.2 The Jacknife 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:

$$\widehat{Bias}_{\text{jack}}(T_n) = (n-1)\left[\frac{1}{n}\sum_{i=1}^{n} T_{n-1}^{(i)} - T_n\right]$$
$$\widehat{Bias}_{\text{boot}}(T_n) = \left[\frac{1}{B}\sum_{b=1}^{B} T_n^{*(b)} - T_n\right]$$

- Variance:

$$\begin{aligned}
\widehat{Var}_{\text{jack}}(T_n) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}(T_{n,i}-\widehat{T}_n)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\Big([nT_n-(n-1)T_{n-1}^{(i)}] \\
&\qquad -[nT_n-\frac{n-1}{n}\sum_{j=1}^{n}T_{n-1}^{(j)}]\Big)^2 \\
&= (n-1)\left[\frac{1}{n}\sum_{i=1}^{n}(T_{n-1}^{(i)}-\frac{1}{n}\sum_{j=1}^{n}T_{n-1}^{(j)})^2\right] \\
\widehat{Var}_{\text{boot}}(T_n) &= \frac{1}{B}\sum_{b=1}^{B}(T_n^{*(b)}-\frac{1}{B}\sum_{b=1}^{B}T_n^{*(b)})^2
\end{aligned}$$

– Both jacknife formulas have an *inflation factor:* $n-1$

- this is needed because the jacknife deviations are smaller than the bootstrap deviations
- the jacknife resamples resemble much more to $\mathcal{X}$ than $\mathcal{X}^*$ do.

• **Comparison of practical performances**

– Since the jacknife uses only n specific resamples from $\mathcal{X}$, whereas the bootstrap makes use of much more information on the data, we expect better performance of the bootstrap.

– But the jacknife is easier to compute (for, say, $n \leq 200$).

– In fact, the jacknife can be viewed as an approximation of the bootstrap. It depends on the statistics.

- An estimator T_n of θ is **linear** if

$$T_n = \frac{1}{n} \sum_{i=1}^{n} \alpha(X_i),$$

for some function α (example: sample mean).

- An estimator T_n of θ 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 α and β (example: sample variance).

- **Estimation of the variance**

– If T_n is linear:

- One can show:

$$\widehat{Var}_{\text{jack}}(T_n) = \frac{n}{n-1} \widehat{Var}_{\text{boot}}(T_n)$$

where $\widehat{Var}_{\text{boot}}(T_n)$ is the true bootstrap variance (computed with the ideal $B = \infty$)

- Example: sample mean, $T_n = \bar{X}_n$, we have

$$\widehat{Var}_{\text{jack}}(T_n) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$$\widehat{Var}_{\text{boot}}(T_n) = \frac{1}{n^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

– 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 jacknife resamples $\mathcal{X}_{(i)}$ (system of n equations in n unknowns $\alpha_i = \alpha(X_i)$).
- Then

$$\widehat{Var}_{\text{jack}}(T_n) = \frac{n}{n-1}\widehat{Var}_{\text{boot}}(T_{\text{lin},n})$$

- The quality of the jacknife estimator depends on the quality of the linear approximation (if T_n highly nonlinear, the jacknife could be inefficient).

- **Estimation of the bias**

– If T_n is quadratic:

- It can be shown that

$$\widehat{Bias}_{\text{jack}}(T_n) = \frac{n}{n-1}\widehat{Bias}_{\text{boot}}(T_n)$$

- Example: plug-in variance $T_n = (1/n)\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$\begin{aligned}\widehat{Bias}_{\text{jack}}(T_n) &= T_n - \widehat{T}_n \\ &= (\frac{1}{n} - \frac{1}{n-1})\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ &= -\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \\ \widehat{Bias}_{\text{boot}}(T_n) &= -\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\end{aligned}$$

– If T_n is not quadratic:

$$\widehat{Bias}_{\text{jack}}(T_n) = \frac{n}{n-1}\widehat{Bias}_{\text{boot}}(T_{\text{quad},n}),$$

where $T_{\text{quad},n}$ is a quadratic approximation of T_n that equals T_n for the jacknife samples.

• **Example:** Monte-Carlo comparison of jacknife 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 jacknife and the bootstrap estimates ($B = 1000$) of **standard error** of 3 statistics:

$$\begin{aligned} T_1 &= \bar{X}_1 \\ T_2 &= r_{12} = \frac{S_{12}}{S_1\, S_2} \\ T_3 &= \left(\bar{X}_1\right)^2 \end{aligned}$$

– 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 jacknife estimates for the nonlinear statistics T_2 and T_3.

– Note that the true value for $Std(T_1) = \frac{\sigma}{\sqrt{n}} = 0.3162$.

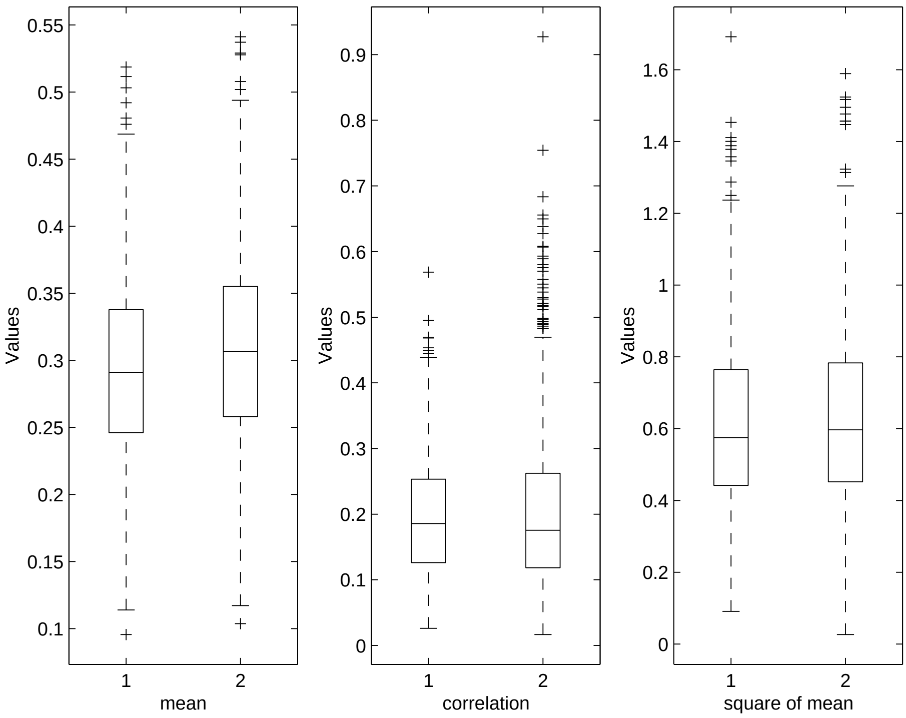

Figure 7.1: *Bootstrap (left) and Jacknife (right) estimates of the standard deviations of 3 statistics, samples of size* $n = 10$ *from a bivarite normal. Left panel* $T_1 = \bar{X}_1$*, middle panel* $T_2 = r_{12}$ *and right panel is* $T_3 = \left(\bar{X}_1\right)^2$*.*

7.3 The Jacknife and the Delta method

• The jacknife can be used to approximate the empirical influence function evaluated at the observed x_i.

– Let $T_n = T(\mathcal{X}) = \theta(F_n)$ be a plug-in estimator of $\theta(F)$. The influence function of θ is:

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

– The nonparametric delta method provides:

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

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

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

– A numerical approximation to the derivative $L_\theta(x_i; F_n)$ is obtained through

$$L_\theta(x_i; F_n) = \frac{\theta\left((1-\varepsilon)F_n + \varepsilon H_{x_i}\right) - \theta(F_n)}{\varepsilon}$$

where $\varepsilon \to 0$ when $n \to \infty$.

– It can be shown that when chosing $\varepsilon = -(n-1)^{-1}$, we obtain the jacknife approximation of $L_\theta(x_i; F_n)$:

$$L_\theta(x_i; F_n) \approx (n-1)\left(T_n - T_{n-1}^{(i)}\right),$$

so we have

$$var_L(F_n) \approx \frac{n-1}{n^2} \sum_{i=1}^{n} \left(T_n - T_{n-1}^{(i)}\right)^2.$$

Chapter 8

The Smoothed Bootstrap

• For the nonparametic 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

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} I\left(x - x_i\right),$$

where $I(\cdot)$ is a dirac-type function puting 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: $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2),$ 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.

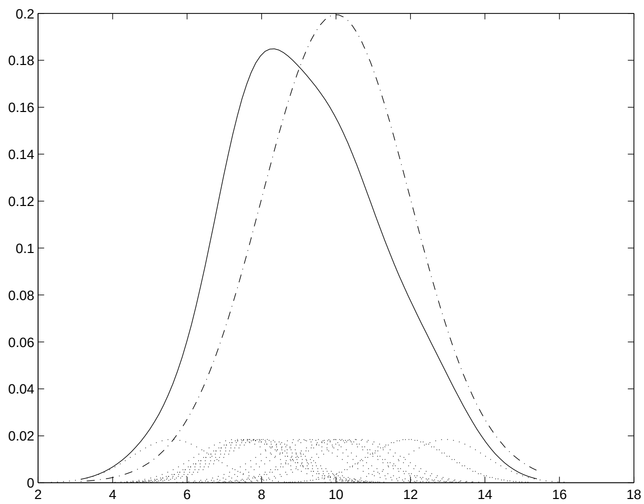

Figure 8.1: *Kernel density estimator: the true density is the dash-dotted line, the kernel estimate is the solid line. The small gaussian densities on the bottom are the kernel functions centered on the n data points and scaled by n: they are added to produce the estimator.*

- **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: $h_R = 1.06 \min\left(\hat{\sigma}, \hat{R}/1.349\right)\, n^{-1/5}$, 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 squarred error (estimation obtain by a leave-on-out technique).

8.2 The bootstrap algorithm

• The idea: resampling from $\widehat{F}_h$ in place of F_n where $\widehat{F}_h$ 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 $\widehat{F}_h$: $X_{i,s}^* \sim \widehat{F}_h$.

– 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 $E(\bar{X}^*)$ 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).

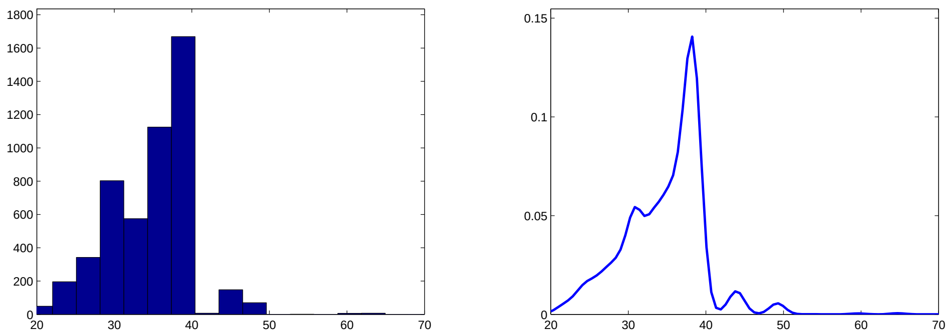

Figure 8.2: *Monte-Carlo distribution of* $\hat{q}^*_{0.95}$ *with naive bootstrap. Here* $B = 5000$.

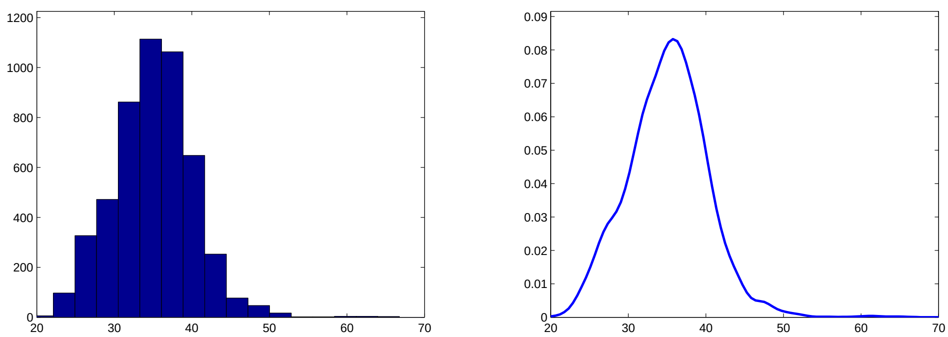

Figure 8.3: *Monte-Carlo distribution of* $\hat{q}^*_{0.95}$ *with smoothed bootstrap with* $h = 3.8761$. *Here* $B = 5000$.

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:

$$T = \min_{h>0}\{h \,|\, \hat{f}_h(x) \text{ is unimodal}\}$$

- We reject if T is too large. Let $t = T_{observed}$.

- The p-value is

$$p - \text{value } = Prob(T \geq t \,|\, H_0)$$

where H_0 is the unimodality of $f(x)$.

– The algorithm:

-[1] we generate B samples $\mathcal{X}^{*,b}_{s,corr}$, $b = 1, \ldots, B$ **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 easely 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 γ_k: for all t, k we have

$$
\begin{aligned}
E(X_t) &= 0 \\
Cov(X_t, X_{t+k}) &= \gamma_k.
\end{aligned}
$$

– 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:

$$\varepsilon_t \sim (0, \sigma^2), \quad \text{independent}$$

  * 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

$$X_t = \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t + \sum_{k=1}^{q} \beta_k \varepsilon_{t-k},$$

    $\{\varepsilon_t\}$ 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 ε_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 $X_0^*, \ldots, X_n^*$:

$$\begin{aligned} X_0^* &= \varepsilon_0^* \\ X_t^* &= \hat{\alpha} X_{t-1}^* + \varepsilon_t^*, \quad t = 1, \ldots, n \end{aligned}$$

  - 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,...), prediction (see Thombs and Schucany, 1990),...

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 lenght ℓ. So we suppose $n = b\ell$.
- Define the blocks y_j:

$$
\begin{aligned}
y_1 &= (x_1, \ldots, x_\ell) \\
y_2 &= (x_{\ell+1}, \ldots, x_{2\ell}) \\
&\ldots \\
y_b &= (x_{(b-1)\ell+1}, \ldots, x_{b\ell})
\end{aligned}
$$

- Take a bootstrap sample of b blocks drawn independently with replacement from the set $(y_1, \ldots, y_b)$.

  - This produce a boostrap sample $(y_1^*, \ldots, y_b^*)$ which gives a bootstrap series of lenght n.
- The idea is that if ℓ 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: ℓ 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 ℓ, 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 easely 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.

$$Y_i = \begin{cases} 0 & \text{if } Y_i^* < 0 \\ 1 & \text{if } Y_i^* \geq 0 \end{cases} \qquad \text{where } Y_i^* | X_i = x_i \sim N(x_i'\beta, 1)$$

    - Equivalently

$$\begin{aligned} P(Y_i = 1 | X_i = x_i) &= \Phi(x_i'\beta, 1) \\ f(y_i | x_i) &= (\Phi(x_i'\beta, 1))^{y_i} \; (1 - \Phi(x_i'\beta, 1))^{1-y_i} \; I_{\{0,1\}}(y_i) \end{aligned}$$

    - 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 η are $\eta_k = x'\beta_k$, $k = 1, \ldots, p$ for some explanatory variables x and some vector β_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 $\mathcal{S}$, for Y is the set of rows $\mathcal{S}_j$ of the identity matrix I_p: $\mathcal{S}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 in the jth column.

- The probability function for Y is now written as a $p - 1$ dimensional integral:

$$\begin{aligned} P(Y = \mathcal{S}_j) &= P(Y_j = 1, Y_k = 0, k \neq j) \\ &= P(Y_j^* - Y_k^* \geq 0, k = 1, \ldots, p, k \neq j) \end{aligned}$$

  * 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

$$f(y; \theta) = \begin{cases} \prod_{j=1}^{p} \Phi_{p-1}(-\mu_j, \Omega_j)^{y_j} & \text{for } y \in \mathcal{S} \\ 0 & \text{otherwise.} \end{cases}$$

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 $y \in \mathcal{S}$, by the empirical Monte-Carlo proportions

- For instance, for $y_i = \mathcal{S}_j$ we have:

$$\begin{aligned}
\widehat{P}(Y = \mathcal{S}_j) &= \hat{f}(y_i; \eta, \Omega) \\
&= \widehat{P}(Y_j^{*(b)} - Y_k^{*(b)} \geq 0, k = 1, \ldots, p, k \neq j) \\
&= \frac{\#\{Y_j^{*(b)} - Y_k^{*(b)} \geq 0, k = 1, \ldots, p, k \neq j\}}{B}.
\end{aligned}$$

- Then we can compute the likelihood function for any values of η, Ω

$$\ln L(\eta, \Omega; y_1, \ldots, y_n) = \sum_{i=1}^{n} \ln \hat{f}(y_i; \eta, \Omega)$$

– 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 that the bootstrap ideas...and the computations are really dependent on the specific particular problem analyzed.

• Some references: Halivassiliou and Ruud (1994), Gouriéroux and Monfort (1995), McFadden (1989).

• **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**
Davison and Hinkley (1997), Efron and Tibshirani (1993), Politis, Romano and Wolf (1999)

• **Introductory surveys**
Hall (1994), Horowitz (1997, 2001, 2003), Politis(1995), Davison, Hinkley and Young (2003)

• **Historical references**
Quenouille (1949), Efron (1982)

• **Theoretical references**
Hall(1992), Beran and Ducharme (1991), Bickel and Freedman (1981), Hall (1988), Mammen (1992), Shao and Tu (1995)

• **Various Applications**

– Linear models: Flachaire (1998), Freedman (1981), Mammen (1993), Stine (1985)

– Time series: Hall, Horowitz and Jing (1995), Härdle, Horowitz and Kreiss (2003), Politis and White (2004)

– Nonparametric regression: Härdle and Bowman (1988), Härdle and Marron (1991)

– Panel data: Hall, Härdle and Simar (1995)

– Frontier models: Hall and Simar (2002), Simar and Wilson (1998, 1999, 2000a, 2000b), Cazals, Florens and Simar (2002), Kneip, Simar and Wilson (2007), Daraio and Simar (2007)

– Smoothed bootstrap: Silverman and Young (1987)

– Bagging: Breiman (1996), Climov, Delecroix and Simar (2001). Simar and Wilson (2007)

– Subsampling: Bertail, Politis and Romano (1999), Bertail, Haefke, Politis and White (2004), Politis, Romano and Wolf (1999, 2001), Bickel and Sakov (2005)

– See the list below, ...

Bibliography

[1] Beran, R. (1986), Discussion of “Jacknife bootstrap and other resampling methods in regression analysis” by C.F.J. Wu, *Annals of Statistics*, 14, 1295–1298.

[2] Beran,R. and G.R. Ducharme (1991). *Asymptotic Theory for the Bootstrap Methods in Statistics*, Les Publications CRM, Centre de Recherches Mathématiques, Université de Montreal, Montreal, Canada.

[3] Bertail, P, C. Haefke, D.N. Politis and H. White (2004), Subsampling the distribution of diverging statistics with applications to finance, *Journal of Econometrics*, 120, 295–326.

[4] Bertail, D.N. Politis and J. Romano (1999), On subsamplinf estimators with unknown rate of convergence, *Journal of the American Statistical Association*, 94, 569–579.

[5] Bickel, P. J. and Freedman, D.A. (1981). Some Asymptotic Theory for the Bootstrap, *The Annals of Statistics*, Vol 9, 6, 1196-1217.

[6] Bickel, P.J. and A. Sakov (2005), On the choice of m in the m Out of n bootstrap and its application to confidence bounds for extreme percentiles, manuscript, forthcoming in ?*Statistica Sinica*?.

[7] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26, 123–140.

[8] Cazals, C. Florens, J.P. and L. Simar (2002) : Nonparametric Frontier Estimation : a Robust Approach, *Journal of Econometrics*, 106, 1–25.

[9] Climov, D., M. Delecroix and L. Simar (2002), Semiparametric Estimation in Single index Poisson Regression: a practical approach, *Journal of Applied Statistics*, vol 29, 7, 1047–1070.

[10] Daraio, C. and L. Simar (2007), *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*, Springer-Verlag, New-York.

[11] Davidson, R. and E. Flachaire (2001), The Wild Bootstrap, Tamed at Last, manuscript, London School of Economics and GREQAM, Marseille.

[12] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*, Cambridge University Press.

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

[14] Efron, B. (1982). *The Jacknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

[15] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[16] Flachaire, E. (1998). A better way to bootstrap pairs. Manuscript, GREQAM, Marseille, France.

[17] Freedman, D.A. (1981). Bootstrapping regression models, *The Annals of Statistics*, Vol 9, 6, 1218-1228.

[18] Gibbons, J.D. (1971),*Non Parametric Statistical Inference*, Mc Graw Hill, New-York.

[19] Gouriéroux, Ch. and A. Montfort (1995), *Simulation Based Econometric Methods*, CORE Lecture Series, CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

[20] Hajivassiliou, V.A. and P.A. Ruud (1994). Classical Estimation Methods for LDV Models Using Simulation, in *Handbook of Econometrics*, Vol IV, 2383–2441. R.F. Engle and D.L. McFadden, eds., Elsevier Sciences B.V.

[21] Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *The Annals of Statistics*, Vol 16, 3, 927-953.

[22] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

[23] Hall, P. (1994). Methodology and Theory for the Bootstrap, in *Handbook of Econometrics*, Vol IV, 2341–2381. R.F. Engle and D.L. McFadden, eds., Elsevier Sciences B.V.

[24] Hall, P., Härdle, W. and L. Simar (1993), "On the Inconsistency of Bootstrap Distribution Estimators", *Computational Statistics and Data Analysis*, 16, 11-18.

[25] Hall, P., Härdle, W. and L. Simar (1995), "Iterated Bootstrap with Application to Frontier Models", *Journal of Productivity Analysis*, 6, p63-76.

[26] Hall, P., J.L. Horowitz and B.Y. Jing (1995). On blocking rules for the bootstrap with dpendent data, *Biometrika*, 82, 561-574.

[27] Hall, P. and L. Simar (2002), Estimating a Changepoint, Boundary or Frontier in the Presence of Observation Error, *Journal of the American Statistical Association*, 97, 523–534.

[28] Härdle, W. and A.W. Bowman (1988). Bootstrapping in nonparametric regression: local adaptative smoothing and confidence bands. *Journal of the American Statistical Association*, 83, 102-110.

[29] Härdle, W. and J.S. Marron (1991). Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19, 778-796.

[30] Härdle, W., Horowitz, J.L. and J.P. Kreiss (2003) A Review of the Bootstrap for Time Series. *International Statistical Review*, 71, 435-459.

[31] Härdle, W. and L. Simar (2003), *Applied Multivariate Statistical Analysis*, Springer-Verlag, Berlin.

[32] Horowitz, J.L. (1997), Bootstrap methods in econometrics: Theory and numerical performance, in *Advances in Economics and Econometrics : Theory and Application*, Vol 3, 188-222. D.M. Kreps and K.F. Wallis, eds., Cambridge University Press.

[33] Horowitz, J.L. (2001), The Bootstrap, in *Handbook of Econometrics*, vol 5, Eds J.J Heckman and E.E Leamer, Amsterdam, North-Holland publishing Company.

[34] Horowitz, J.L. (2003), The bootstrap in econometrics, *Statistical Science*, 18, 211–218.

[35] Kneip, A, L. Simar and P.W. Wilson (2007), Asymptotics for DEA Estimators in Nonparametric Frontier Models, Discussion paper 0317, Institut de Statistique, UCL, accepted in *Economteric Theory.*

[36] Lahiri, S.N. (1999), Theoretical comparisons of block bootstrap methods, *Annals of Statistics*, 27, 386–404.

[37] Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions, *Econometric Reviews*, 15(3), 261–274.

[38] Li, Q. (1999). Nonparametric testing the similarity of two unknown density functions: local power and bootstrap analysis, *Nonparametric Statistics*, 11, 189–213.

[39] Liu, R.Y. (1988). Bootstrap procedures under some non i.i.d. models, *Annals of Statistics*, 16, 1696–1708.

[40] MacKinnon, J.G. and H. White (1985). Some Heteroskedasticity consistent covariance matrix estimators with improved finite sample properties, *Journal of Econometrics*, 29, 305–325.

[41] McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica*, Vol 57, 5, 995–1026.

[42] Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulation*, Springer-Verlag, New York.

[43] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensionnal linear models *The Annals of Statistics*, Vol 21, 1, 255-285.

[44] Politis, D.N. (1995). A Primer on Bootstrap Methods in Satistics, Technical Report #95-19, Department of Statistics, Purdue University.

[45] Politis, D.N. and J.P. Romano (1994), The stationary bootstrap, *Journal of the American Statistical Association*, 89, 1303–1313..

[46] Politis, D.N., J.P. Romano and M. Wolf (1999), *Subsampling*, Springer, New-York.

[47] Politis, D.N., J.P. Romano and M. Wolf (2001), On the asymptotic Theory of subsampling, *Statistica Sinica*, 11, 1105–1124.

[48] Politis, D.N. and H. White (2004), Automatic Block-length selection for the dependent bootstrap, *Econometric Review*, 23, 53–70.

[49] Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, B*, 11, 68–84.

[50] Serfling, R.S. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New-York.

[51] Shao, J. and Tu, D. (1995). *The Jacknife and Bootstrap*, Springer-Verlag, New York.

[52] Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[53] Silverman, B.W. and G.A. Young (1987), The bootstrap: smooth or not to smooth?, *Biometrika* 74, 469-479.

[54] Simar, L. and P. Wilson (1998), Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* 44, 49–61.

[55] Simar, L. and P. Wilson (1999) : "Estimating and Bootstrapping Malmquist Indices", *European Journal of Operational Research*, 115, 459-471.

[56] Simar L. and P. Wilson (2000a) : "Statistical Inference in Nonparametric Frontier Models : The State of the Art", *Journal of Productivity Analysis*, 13, 49-78.

[57] Simar L. and P. Wilson (2000b) : "A General Methodology for Bootstrapping in Nonparametric Frontier Models", *Journal of Applied Statistics*, Vol 27, 6, 779-802.

[58] Simar, L. and P.W. Wilson (2007), Estimation and Inference in Cross-Sectional Stochastic Frontier Models. Discussion paper 0524, revised version 2007, Institut de Statistique, UCL.

[59] Stine, R. A. (1985), Bootstrap prediction intervals for regression, *Journal of the American Statistical Association* 80, 1026-1031.

[60] Thombs, L.A. and W.R. Schucany (1990), Bootstrap prediction intervals for autoregression, *Journal of the American Statistical Association* 85, 486–493.

[61] Tukey, J.W. (1958) Bias and confidence in not quite large samples. (Abstract.), *Annals of Mathematical Statistics*, 29, 614.

Appendix A: Gamma and related distributions

- **Some reminders and notations**

  – Gamma distribution: $X \sim \Gamma(\alpha, \beta)$

$$
\begin{aligned}
E(X) &= \alpha\beta \\
V(X) &= \alpha\beta^2 \\
f(x) &= \frac{x^{\alpha-1}\exp(-x/\beta)}{\beta^{\alpha}\Gamma(\alpha)}
\end{aligned}
$$

  – Exponential distribution: $X \sim \mathrm{Expo}(\theta) \equiv \Gamma(1, \theta)$

$$
\begin{aligned}
E(X) &= \theta \\
V(X) &= \theta^2 \\
f(x) &= \frac{\exp(-x/\theta)}{\theta}
\end{aligned}
$$

  – Chi-square distribution: $X \sim \chi^2_\nu \equiv \Gamma(\nu/2, 2)$.

$$
\begin{aligned}
E(X) &= \nu \\
V(X) &= 2\nu
\end{aligned}
$$

- **Normal sampling**

  – $X_i \sim IN(\mu, \sigma^2)$, $i = 1, \ldots, n$. We have:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$
$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1} \equiv \Gamma(\frac{n-1}{2}, 2\frac{\sigma^2}{n-1})$$

- **Exponential sampling**

  – $X_i \sim \text{Expo}(\mu)$, $i = 1, \ldots, n$. We have:

$$\bar{X} \sim \Gamma(n, \frac{\mu}{n})$$
$$\frac{\bar{X}}{\mu} \sim \Gamma(n, \frac{1}{n}).$$