Bootstrap and Jackknife

- in Statistics, we deal with the difficulty of finding the distribution / standard errors of uncommon test statistics
- Bootstrap and Jackknife are general recipes to attack this problem
- Bootstrap offers much more, as we will find out

Plug-in Estimators I

- we have independent observations $\mathcal{X}_n := (X_1, X_2, \ldots, X_n)$ from an unknown density $f(\cdot)$ with c.d.f. $F(\cdot)$
- want to estimate an interesting feature of the distribution $\theta = T(F)$, where $T(\cdot)$ is a function of the c.d.f., e.g.,
  - mean: if $\theta = E(X_1) = \int x f(x)\, dx = \int x\, dF(x)$ then $T(G) = \int x\, dG(x)$, where $G(\cdot)$ is any c.d.f.
  - median: if $\theta = F^{-1}(0.5)$ then $T(F) = G^{-1}(0.5)$, where $G(\cdot)$ is any c.d.f.
- say $\widehat{F}_n(\cdot)$ is the empirical c.d.f. of the data $\mathcal{X}_n$, i.e., $\widehat{F}_n(\cdot)$ puts mass $1/n$ on each of the data points $X_i,\ i = 1, 2, \ldots, n$
- then one could use the *plug-in* estimator of θ: $\widehat{\theta}_n := T(\widehat{F}_n) = t(\mathcal{X}_n)$, e.g.,
  - mean: if $T(F) = \int x\, dG(x)$ then $T(\widehat{F}_n) = \int x\, d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}_n = t(\mathcal{X}_n)$, the sample mean
  - median: if $T(G) = G^{-1}(0.5)$ then $T(\widehat{F}_n) = \widehat{F}_n^{-1}(0.5) = t(\mathcal{X}_n)$, the sample median

Plug-in Estimators II

- plug-in estimator of the variance:

$$\theta = E(X_1 - E(X_1))^2 = E(X_1^2) - (E(X_1))^2$$

$$= \int x^2 \, dF(x) - \left\{ \int x \, dF(x) \right\}^2 = T(F) \quad \text{and hence}$$

$$T(\widehat{F}_n) = \int x^2 \, d\widehat{F}_n(x) - \left\{ \int x \, d\widehat{F}_n(x) \right\}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i \right\}^2$$

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

- note here $E(T(\widehat{F}_n)) = \frac{n-1}{n} E(s_n^2) = \frac{n-1}{n} \theta \neq \theta$, i.e., the plug-in estimator is biased

Bootstrap Principle

- inference on θ is usually based on the distribution of $T(\widehat{F}_n) = t(\mathcal{X}_n)$ or of $R(\mathcal{X}_n, F) := \frac{T(\widehat{F}_n) - T(F)}{S(\widehat{F}_n)}$, where $S(\cdot)$ is a functional, which estimates $Var(T(\widehat{F}_n))$ or for other forms of $R(\mathcal{X}_n, F)$
- both of the above distributions may be intractable and it also may depend on the unknown $F(\cdot)$
- bootstrap provides a way out using random *bootstrap samples* or *pseudo-data sets* denoted by $\mathcal{X}_n^\star := (X_1^\star, X_2^\star, \ldots, X_n^\star)$, we'll see how to generate these in a bit
- let $\widehat{F}_n^\star(\cdot)$ be the empirical c.d.f. of the data $\mathcal{X}_n^\star$, i.e., $\widehat{F}_n^\star(\cdot)$ puts mass $1/n$ on each of the data points $X_i^\star,\ i = 1, 2, \ldots, n$
- since $\mathcal{X}_n^\star$ is randomly generated, $\widehat{F}_n^\star(\cdot)$ is a random variable
- now the *bootstrap principle* says that approximate the distribution of
  - $T(\widehat{F}_n) = t(\mathcal{X}_n)$ by the bootstrap distribution of $T(\widehat{F}_n^\star) = t(\mathcal{X}_n^\star)$
  - $R(\mathcal{X}_n, F)$ by the bootstrap distribution of $R(\mathcal{X}_n^\star, \widehat{F}_n)$

Bootstrap Sample or Pseudo-data Set

- **Non-parametric Bootstrap**:

$$\text{data world:} \quad \mathcal{X}_n := (X_1, X_2, \ldots, X_n),\ X_i \overset{i.i.d.}{\sim} F(\cdot),\ i = 1, 2, \ldots, n$$

$$\text{bootstrap world:} \quad \mathcal{X}_n^\star := (X_1^\star, X_2^\star, \ldots, X_n^\star),\ X_i^\star \overset{i.i.d.}{\sim} \widehat{F}_n(\cdot),\ i = 1, 2, \ldots, n$$

  here the pseudo-data set is a simple random sample of size n with replacement from $\mathcal{X}_n := (X_1, X_2, \ldots, X_n)$

- **Parametric Bootstrap**: let θ and $\widehat{\theta}_n$ be a parameter and it's reasonable estimator below (may not necessarily be the plug-in estimator!)

$$\text{data world:} \quad \mathcal{X}_n := (X_1, X_2, \ldots, X_n),\ X_i \overset{i.i.d.}{\sim} F(\cdot, \theta),\ i = 1, 2, \ldots, n$$

$$\text{bootstrap world:} \quad \mathcal{X}_n^\star := (X_1^\star, X_2^\star, \ldots, X_n^\star),\ X_i^\star \overset{i.i.d.}{\sim} F(\cdot, \widehat{\theta}_n),\ i = 1, 2, \ldots, n$$

  here the pseudo-data set is a random sample of size n from $F(\cdot, \widehat{\theta}_n)$

Bootstrap Distribution

- note that both in non-parametric and in parametric bootstrap, the bootstrap samples are generated "using" the observed data $\mathcal{X}_n$, i.e., conditioned on $\mathcal{X}_n$
- for non-parametric bootstrap the number of possible bootstrap samples is $n^n (\approx \infty, \text{ for large } n)$
- for parametric bootstrap the number of possible bootstrap samples is ∞
- since it's not possible or efficient to consider all possible bootstrap samples we work with B-many bootstrap samples
  $\mathcal{X}^\star_{n,b} = \{X^\star_{1,b}, X^\star_{2,b}, \ldots, X^\star_{n,b}\},\ b = 1, 2, \ldots, B$
- then the bootstrap distribution of $T(\widehat{F}^\star_n) = t(\mathcal{X}^\star_n)$ and of $R(\mathcal{X}^\star_n, \widehat{F}_n)$ can be approximated by the following empirical distributions of

$$\widehat{G}^\star_{B,T}(\cdot) \text{ is the c.d.f. of } \left\{ T(\widehat{F}^\star_{n,b}) = t(\mathcal{X}^\star_{n,b}) \mid b = 1, 2, \ldots, B \right\} \quad \text{and}$$

$$\widehat{G}^\star_{B,R}(\cdot) \text{ is the c.d.f. of } \left\{ R(\mathcal{X}^\star_{n,b}, \widehat{F}_n) \mid b = 1, 2, \ldots, B \right\}$$

Bootstrap Standard Error

- let us consider the problem of estimating $\theta = T(F)$ by the plug-in estimator $T(\widehat{F}_n) = t(\mathcal{X}_n)$, say
- interested in the standard error of the estimator $T(\widehat{F}_n)$, i.e., we want to estimate $\sqrt{Var_F(T(\widehat{F}_n))} = \sqrt{Var_F(t(\mathcal{X}_n))}$
- use the sample standard deviation of the c.d.f. $\widehat{G}^{\star}_{B,T}(\cdot)$ or of the (bootstrap sample) evaluations $\{t(\mathcal{X}^{\star}_{n,b}) \mid b = 1, 2, \ldots, B\}$ as an estimator of $\sqrt{Var_F(T(\widehat{F}_n))} = \sqrt{Var_F(t(\mathcal{X}_n))}$:

$$\text{se}_{boot,B}(t) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(t(\mathcal{X}^{\star}_{n,b}) - \bar{t}^{\star}_{B} \right)^2 \right\}^{1/2} \quad \text{where } \bar{t}^{\star}_{B} = \frac{1}{B} \sum_{b=1}^{B} t(\mathcal{X}^{\star}_{n,b})$$

Non-parametric Bootstrap: Regression I

- the model: $Y_i = X_i^T \beta + \epsilon_i,\ i = 1, 2, \ldots, n$
- want the standard error of $\widehat{\beta}_{OLS}$
- define $Z_i := (X_i^T, Y_i),\ i = 1, 2, \ldots, n$
- generate B-many pseudo-data sets $\mathcal{Z}^\star_{n,b} = \{Z^\star_{1,b}, Z^\star_{2,b}, \ldots, Z^\star_{n,b}\}$ by sampling with replacement from $\{Z_1, Z_2, \ldots, Z_n\}$ randomly
- for $b = 1, 2, \ldots, B$, regress (OLS) $\{Y^\star_{i,b} \mid i = 1, 2, \ldots, n\}$ on $\{X^{\star T}_{i,b} \mid i = 1, 2, \ldots, n\}$ to get $\widehat{\beta}^\star_b = t(\mathcal{Z}^\star_{n,b})$, say
- now compute bootstrap standard error $\mathrm{se}_{boot,B}(t)$, to get the standard error of $\widehat{\beta}_{OLS}$
- this is also called *paired bootstrapping*

Non-parametric Bootstrap: Regression II

- the model: $Y_i = X_i^T \beta + \epsilon_i,\ i = 1, 2, \ldots, n$
- want standard error of $\widehat{\beta}_{OLS}$
- the (OLS) estimated model: $Y_i = X_i^T \widehat{\beta}_{OLS} + \widehat{\epsilon}_i,\ i = 1, 2, \ldots, n$
- generate pseudo-data sets $\mathcal{E}^\star_{n,b} = \{\widehat{\epsilon}^\star_{1,b}, \widehat{\epsilon}^\star_{2,b}, \ldots, \widehat{\epsilon}^\star_{n,b}\}$ by sampling with replacement from $\{\widehat{\epsilon}_1, \widehat{\epsilon}_2, \ldots, \widehat{\epsilon}_n\}$ randomly
- now define $Z^\star_i = (Y^\star_i := X_i^T \widehat{\beta}_{OLS} + \widehat{\epsilon}^\star_i, X_i)$ and form $\mathcal{Z}^\star_{n,b} = \{Z^\star_{1,b}, Z^\star_{2,b}, \ldots, Z^\star_{n,b}\}$
- for $b = 1, 2, \ldots, B$, regress (OLS) $\{Y^\star_{i,b} \mid i = 1, 2, \ldots, n\}$ on $\{X_i \mid i = 1, 2, \ldots, n\}$ to get $\widehat{\beta}^\star_b = t(\mathcal{Z}^\star_{n,b})$, say
- now compute bootstrap standard error $\mathrm{se}_{boot,B}(t)$, to get the standard error of $\widehat{\beta}_{OLS}$
- this is also called *bootstrapping the residuals*

Parametric Bootstrap: Gamma model

- the model: $Y_i \overset{i.i.d.}{\sim} \Gamma(\alpha, 1),\ i = 1, 2, \ldots, n$
- want the standard error of $\widehat{\alpha}_{MLE}$
- note $\widehat{\alpha}_{MLE}$ is gotten by numerically solving the "score equation"
- generate B-many pseudo-data sets $\mathcal{Y}^{\star}_{n,b} = \{Y^{\star}_{1,b}, Y^{\star}_{2,b}, \ldots, Y^{\star}_{n,b}\}$ by drawing independently from the distribution $\Gamma(\widehat{\alpha}_{MLE}, 1)$
- now compute bootstrap standard error $\mathrm{se}_{boot,B}(t)$, to get the standard error of $\widehat{\alpha}_{MLE}$
- this is preferable than the non-parametric version if we strongly believe that the data generation process, which is a $\Gamma(\cdot, 1)$ distribution here

Bootstrap Estimate Of Bias

- let us consider the problem of estimating $\theta = T(F)$ by the plug-in estimator $T(\widehat{F}_n) = t(\mathcal{X}_n) = \widehat{\theta}_n$
- want to estimate the bias in estimation:
  $E_F(\widehat{\theta}_n) - \theta = E_F(t(\mathcal{X}_n)) - \theta = E_F(T(\widehat{F}_n)) - T(F) = R(\mathcal{X}_n, F)$, say
- we estimate it by:
  $R(\mathcal{X}_n^\star, \widehat{F}_n) = E_{\widehat{F}_n}(T(\widehat{F}_n^\star)) - T(\widehat{F}_n) = E_{\widehat{F}_n}(t(\mathcal{X}_n^\star)) - t(\mathcal{X}_n) = E_{\widehat{F}_n}(\widehat{\theta}_n^\star) - \widehat{\theta}_n$
- use the c.d.f. $\widehat{G}_{B,R}^\star(\cdot)$ or the (bootstrap sample) evaluations
  $\{t(\mathcal{X}_{n,b}^\star) = \widehat{\theta}_{n,b}^\star \mid b = 1, 2, \ldots, B\}$ to approximate by $E_{\widehat{F}_n}(T(\widehat{F}_n^\star)) - \widehat{\theta}_n$

$$\widehat{\text{bias}}_{boot,B}(t) := \frac{1}{B}\sum_{b=1}^{B} t(\mathcal{X}_{n,b}^\star) - t(\mathcal{X}_n) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}_{n,b}^\star - \widehat{\theta}_n$$

Jackknife I

- we have independent observations $\mathcal{X}_n := (X_1, X_2, \ldots, X_n)$ from an unknown density $f(\cdot)$ with c.d.f. $F(\cdot)$,
- consider the problem of estimating $\theta = T(F)$ using $\widehat{\theta}_n = t(\mathcal{X}_n)$
- form the "leave-one" data sets:
  $\mathcal{X}_{(-i)} := \{X_j \mid j = 1, 2, \ldots, n, j \neq i\},\ i = 1, 2, \ldots, n$
- define $\widehat{\theta}_{(-i)} = t(\mathcal{X}_{(-i)}),\ i = 1, 2, \ldots, n$ and $\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_{(-i)}$
- then the Jackknife estimate of the bias $E_F(\widehat{\theta}_n) - \theta = E_F(t(\mathcal{X}_n)) - \theta$ is given by:

$$\widehat{\text{bias}}_{jack}(t) = (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_n)$$

Jackknife II

- then the Jackknife estimate of the standard error $\sqrt{Var_F(\widehat{\theta}_n)} = \sqrt{Var_F(t(\mathcal{X}_n))}$ is given by:

$$\text{se}_{jack}(t) = \left\{ \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\theta}_{(-i)} - \widehat{\theta}_{(\cdot)} \right)^2 \right\}^{1/2}$$

- why do the formulas for $\widehat{\text{bias}}_{jack}(t)$ and $\text{se}_{jack}(t)$ make sense?
  - define the *pseudo-values* $\tilde{\theta}_i := n\widehat{\theta}_n - (n-1)\widehat{\theta}_{(-i)},\ i = 1, 2, \ldots, n$ and their mean $\bar{\tilde{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\theta}_i$
  - now note,

$$\widehat{\text{bias}}_{jack}(t) = (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta}_n) = \widehat{\theta}_n - \bar{\tilde{\theta}} \quad \text{and}$$

$$\text{se}_{jack}(t) = \left\{ \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\theta}_{(-i)} - \widehat{\theta}_{(\cdot)} \right)^2 \right\}^{1/2} = \left\{ \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2 \right\}^{1/2} \quad (1)$$

Jackknife III

- note if $\widehat{\theta}_n = \bar{X}_n$, the sample mean then $\tilde{\theta}_i = X_i,\ i = 1, 2, \ldots, n$ and $\bar{\tilde{\theta}} = \bar{X}_n$
- here the above formulas in equation (1) become

$$\widehat{\text{bias}}_{jack}(t) = \widehat{\theta}_n - \bar{\tilde{\theta}} = 0 \quad \text{and}$$

$$\text{se}_{jack}(t) = \left\{ \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\tilde{\theta}_i - \bar{\tilde{\theta}} \right)^2 \right\}^{1/2} = \left\{ \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(X_i - \bar{X}_n \right)^2 \right\}^{1/2}$$

- thus the formulas in equation (1) "does the right thing" in case of $\widehat{\theta}_n = \bar{X}_n$
- hope is they will do a good job for a general statistic $\widehat{\theta}_n$ as well

Bootstrap Confidence Intervals I

- let us consider the problem of estimating $\theta = T(F)$ by the plug-in estimator $T(\widehat{F}_n) = t(\mathcal{X}_n) = \widehat{\theta}_n$, say
- let $R(\mathcal{X}_n, F) := \frac{T(\widehat{F}_n) - T(F)}{S(\widehat{F}_n)}$ be roughly "pivotal" for the problem, (i.e. we expect the distribution of $R(\mathcal{X}_n, F)$ under $F(\cdot)$ to be free of θ)
- want a $100(1-\alpha)\%$ confidence interval for θ
- let we have the c.d.f.s $\widehat{G}^{\star}_{B,T}(\cdot)$ and $\widehat{G}^{\star}_{B,R}(\cdot)$ from B-many pseudo-data sets $\mathcal{X}^{\star}_{n,b} = \{X^{\star}_{1,b}, X^{\star}_{2,b}, \ldots, X^{\star}_{n,b}\}$
- let, in general, $\zeta_{\gamma}(G)$ denote the γ-th percentile of the c.d.f.: $G(\cdot),\ \gamma \in (0,1)$

Bootstrap Confidence Intervals II

- then one could use one of three kinds of confidence intervals:
  - **Percentile interval**: the most basic one:

$$\left[\zeta_{\alpha/2}(\widehat{G}^{\star}_{B,T}), \zeta_{1-\alpha/2}(\widehat{G}^{\star}_{B,T})\right]$$

  - **Accelerated bias-corrected percentile interval or BC_a interval**: percentile based but has better coverage:

$$\left[\zeta_{\beta_1}(\widehat{G}^{\star}_{B,T}), \zeta_{\beta_2}(\widehat{G}^{\star}_{B,T})\right],$$

    where $\beta_1, \beta_2 \in (0,1)$ and have complicated expressions, we'll see later
  - **Bootstrap t interval**: in case we have a rough pivot:

$$\left[T(\widehat{F}_n) - S(\widehat{F}_n)\zeta_{1-\alpha/2}(\widehat{G}^{\star}_{B,R}), T(\widehat{F}_n) - S(\widehat{F}_n)\zeta_{\alpha/2}(\widehat{G}^{\star}_{B,R})\right]$$

Percentile Interval I

- $H(\cdot)$ is a continuous and symmetric (around 0) distribution
- let $\psi(\cdot)$ be a continuous, strictly increasing transformation such that $\psi(\widehat{\theta}_n) - \psi(\theta) \sim H(\cdot)$, e.g. $\psi(\cdot)$ could be a normalizing transformation
- then we have, by symmetry of $H(\cdot)$:

$$P\left[\zeta_{\alpha/2}(H) \leq \psi(\widehat{\theta}_n) - \psi(\theta) \leq \zeta_{1-\alpha/2}(H)\right] = 1-\alpha \quad (2)$$

$$\implies P\left[\psi^{-1}(-\zeta_{1-\alpha/2}(H) + \psi(\widehat{\theta}_n)) \leq \theta \leq \psi^{-1}(-\zeta_{\alpha/2}(H) + \psi(\widehat{\theta}_n))\right] = 1-\alpha$$

$$\implies P\left[\psi^{-1}(\zeta_{\alpha/2}(H) + \psi(\widehat{\theta}_n)) \leq \theta \leq \psi^{-1}(\zeta_{1-\alpha/2}(H) + \psi(\widehat{\theta}_n))\right] = 1-\alpha \quad (3)$$

- using the bootstrap principle on equation (2) we have,

$$P^\star\left[\zeta_{\alpha/2}(H) \leq \psi(\widehat{\theta}_n^\star) - \psi(\widehat{\theta}_n) \leq \zeta_{1-\alpha/2}(H)\right] \approx 1-\alpha$$

$$\implies P^\star\left[\psi^{-1}(\zeta_{\alpha/2}(H) + \psi(\widehat{\theta}_n)) \leq \widehat{\theta}_n^\star \leq \psi^{-1}(\zeta_{1-\alpha/2}(H) + \psi(\widehat{\theta}_n))\right] \approx 1-\alpha \quad (4)$$

Percentile Interval II

- compare equations (3), (4) and note that their upper and lower limits coincide
- but from equation (4), we can take

$$\psi^{-1}(\zeta_{\alpha/2}(H) + \psi(\widehat{\theta}_n)) \approx \zeta_{\alpha/2}(\widehat{G}^{\star}_{B,T}) \quad \text{and}$$
$$\psi^{-1}(\zeta_{1-\alpha/2}(H) + \psi(\widehat{\theta}_n)) \approx \zeta_{1-\alpha/2}(\widehat{G}^{\star}_{B,T})$$

- thus the required interval for $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\zeta_{\alpha/2}(\widehat{G}^{\star}_{B,T}), \zeta_{1-\alpha/2}(\widehat{G}^{\star}_{B,T})\right]$$

- note explicit specification of the transformation $\psi(\cdot)$ is not necessary

BC_a Interval I

- let $\psi(\cdot)$ be a continuous, strictly increasing transformation and $a, b \in \mathbb{R}^1$ such that $1 + a\psi(\theta) > 0$ and

$$U := \frac{\psi(\widehat{\theta}_n) - \psi(\theta)}{1 + a\psi(\theta)} + b \sim \mathbf{Normal}_1(0, 1)$$

- note $a = b = 0$ takes us back to the simple percentile method
- with z_γ denote the γ-th percentile of the $\mathbf{Normal}_1(0, 1)$ distribution we have,

$$P\left[z_{\alpha/2} \le U \le z_{1-\alpha/2}\right] = 1 - \alpha \tag{5}$$

$$\implies P\left[k_1(a, b, 1 - \alpha/2, \widehat{\theta}_n) \le \theta \le k_1(a, b, \alpha/2, \widehat{\theta}_n)\right] = 1 - \alpha, \quad \text{where} \tag{6}$$

$$k_1(a, b, \gamma, \widehat{\theta}_n) = \psi^{-1}\left(\psi(\widehat{\theta}_n) + \frac{(b - z_\gamma)[1 + a\psi(\widehat{\theta}_n)]}{1 - a(b - z_\gamma)}\right) \tag{7}$$

BC_a Interval II

- let $U^\star = \frac{\psi(\widehat{\theta}_n^\star) - \psi(\widehat{\theta}_n)}{1 + a\psi(\widehat{\theta}_n)} + b$ and use the bootstrap principle on equation (5) to get,

$$\begin{aligned} & P^\star\left[z_{\alpha/2} \le U^\star \le z_{1-\alpha/2}\right] && \approx 1-\alpha \\ \Longrightarrow\ & P^\star\left[k_2(a,b,1-\alpha/2,\widehat{\theta}_n) \le \widehat{\theta}_n^\star \le k_2(a,b,\alpha/2,\widehat{\theta}_n)\right] && \approx 1-\alpha \qquad (8) \\ & k_2(a,b,\gamma,\widehat{\theta}_n) && = \psi^{-1}\left(\psi(\widehat{\theta}_n) + (z_\gamma - b)[1 + a\psi(\widehat{\theta}_n)]\right) \qquad (9) \end{aligned}$$

- compare equations (6), (8) and note that these equations will read the same if we choose β_1 and β_2 such that

$$\begin{aligned} k_2(a,b,\beta_1,\widehat{\theta}_n) &= k_1(a,b,1-\alpha/2,\widehat{\theta}_n) \quad \text{and} \\ k_2(a,b,\beta_2,\widehat{\theta}_n) &= k_1(a,b,\alpha/2,\widehat{\theta}_n) \end{aligned}$$

BC_a Interval III

- choice of β_1 boils down to (below $\Phi(\cdot)$ is the c.d.f. of the $\mathbf{Normal}_1(0,1)$ distribution):

$$\psi^{-1}\left(\psi(\widehat{\theta}_n) + (z_{\beta_1} - b)[1 + a\psi(\widehat{\theta}_n)]\right) = \psi^{-1}\left(\psi(\widehat{\theta}_n) + \frac{(b - z_{1-\alpha/2})[1 + a\psi(\widehat{\theta}_n)]}{1 - a(b - z_{1-\alpha/2})}\right)$$

$$\Longleftrightarrow z_{\beta_1} - b = \frac{(b - z_{1-\alpha/2})}{1 - a(b - z_{1-\alpha/2})} \qquad \Longleftrightarrow \beta_1 = \Phi\left(b + \frac{b - z_{1-\alpha/2}}{1 - a(b - z_{1-\alpha/2})}\right)$$

- similarly, we get, $\beta_2 = \Phi\left(b + \frac{b - z_{\alpha/2}}{1 - a(b - z_{\alpha/2})}\right)$
- but from equation (8), we can take

$$k_2(a, b, \beta_1, \widehat{\theta}_n) \approx \zeta_{\beta_1}(\widehat{G}^{\star}_{B,T}) \quad \text{and}$$
$$k_2(a, b, \beta_2, \widehat{\theta}_n) \approx \zeta_{\beta_2}(\widehat{G}^{\star}_{B,T})$$

BC_a Interval IV

- thus the required interval for $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\zeta_{\beta_1}(\widehat{G}^{\star}_{B,T}), \zeta_{\beta_2}(\widehat{G}^{\star}_{B,T})\right]$$

- note, both β_1 and β_2 depends on a and b and usually one takes

$$b = \Phi^{-1}\left(\widehat{G}^{\star}_{B,T}(\widehat{\theta}_n)\right) \quad \text{and}$$

$$a = \frac{\sum_{i=1}^{n} \tau_i^3}{6\left(\sum_{i=1}^{n} \tau_i^2\right)^{3/2}} \quad \text{where}$$

$$\tau_i = \widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(-i)} \quad \text{remember jackknife!}$$

- note explicit specification of the transformation $\psi(\cdot)$ is not necessary
- this method just "corrects the percentiles", i.e. uses β_1 and β_2 instead of $1-\alpha_2$ and $\alpha/2$ from the percentile method

Bootstrap t Interval

- let $H(\cdot)$ be c.d.f. of the distribution of $R(\mathcal{X}_n, F) := \frac{T(\widehat{F}_n)-T(F)}{S(\widehat{F}_n)}$, where $\theta = T(F)$
- then we have:

$$\begin{aligned} & P\left[\zeta_{\alpha/2}(H) \leq R(\mathcal{X}_n, F) \leq \zeta_{1-\alpha/2}(H)\right] && = 1-\alpha \\ \Longrightarrow\ & P\left[\widehat{\theta}_n - S(\widehat{F}_n)\zeta_{1-\alpha/2}(H) \leq \theta \leq \widehat{\theta}_n + S(\widehat{F}_n)\zeta_{\alpha/2}(H)\right] && = 1-\alpha \qquad (10) \end{aligned}$$

- using the bootstrap principle on the distribution $H(\cdot)$ we have $\zeta_\gamma(H) \approx \zeta_\gamma(\widehat{G}^\star_{B,R}),\ \gamma \in (0,1)$
- so the required interval from equation (10) is:

$$\left[T(\widehat{F}_n) - S(\widehat{F}_n)\zeta_{1-\alpha/2}(\widehat{G}^\star_{B,R}), T(\widehat{F}_n) - S(\widehat{F}_n)\zeta_{\alpha/2}(\widehat{G}^\star_{B,R})\right]$$

Bootstrap vs. Jackknife I

- the set up (from problem 4 on problem set 5):
  - let we have $\mathcal{X}_n := (X_1, X_2, \ldots, X_n) \overset{i.i.d.}{\sim} f(\cdot)$ with c.d.f. $F(\cdot)$
  - let we have a linear statistic of the form $\widehat{\theta}_n = \mu + \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i)$
- you showed that for this statistic that the variance of the ideal bootstrap estimator $Var_{\widehat{F}_n}(\widehat{\theta}_n^\star)$ and that of the jackknife estimator $Var_{jack}(\widehat{\theta}_n^\star)$ differ by only a factor of $\frac{n-1}{n}$
- let we have a quadratic statistic of the form $\widehat{\theta}_n = \mu + \frac{1}{n}\sum_{1\le i\le n} \alpha(X_i) + \frac{1}{n^2}\sum_{1\le i<j\le n} \beta(X_i, X_j)$
- the ideal bootstrap bias estimate, $E_{\widehat{F}_n}(\widehat{\theta}_n^\star) - \widehat{\theta}_n$ and the jackknife bias estimate, $\widehat{\text{bias}}_{jack}(t)$ differ by only a factor of $\frac{n-1}{n}$

Bootstrap vs. Jackknife II

- consider the problem of estimating $\theta = T(F)$ using a plug-in estimator $\widehat{\theta}_n = t(\mathcal{X}_n) = T(\widehat{F}_n)$, where $\mathcal{X}_n = (X_1, X_2, \ldots, X_n)$
- let $\boldsymbol{P}^\star := (P_1^\star, P_2^\star, \ldots, P_n^\star)^T$ be probability vector: $0 \leq P_i^\star \leq 1, \ \forall i$ and $\sum_{i=1}^n P_i^\star = 1$
- let $F(\boldsymbol{P}^\star)$ be the c.d.f of the p.m.f. which puts mass $P_i^\star$ on $X_i, \ \forall i$
- by convention, denote $T(F(\boldsymbol{P}^\star))$ by $T(\boldsymbol{P}^\star)$
- note, $\widehat{\theta}_n = T(\widehat{F}_n) = T(\boldsymbol{P}^0)$, where $P_i^0 = 1/n, \ \forall i$
- note, the "leave-one" jackknife estimator $\widehat{\theta}_{(-i)} := t(\mathcal{X}_{(-i)}) = T(\boldsymbol{P}_{(-i)})$, where $P_{(-i),j} = 1/(n-1), \ \forall j \neq i$ and $P_{(-i),i} = 0$

Bootstrap vs. Jackknife III

- consider the random variable $\boldsymbol{P}^\star$ such that

$$n \cdot \boldsymbol{P}^\star \sim \textbf{Multinomial}(n, \boldsymbol{P}^0)$$

- recall, a non-parametric bootstrap sample or pseudo-data set $\mathcal{X}_n^\star := (X_1^\star, X_2^\star, \ldots, X_n^\star)$ is generated by $X_i^\star \overset{i.i.d.}{\sim} \widehat{F}_n(\cdot),\ i = 1, 2, \ldots, n$
- the above procedure is equivalent to the following, which uses $\boldsymbol{P}^\star$:

$$\begin{aligned} \text{generate} \quad & \boldsymbol{M}^\star := n \cdot \boldsymbol{P}^\star \sim \textbf{Multinomial}(n, \boldsymbol{P}^0) \\ \text{set} \quad & \mathcal{X}_n^\star = \{M_i^\star \text{ many copies of } X_i\}, \quad \text{why?} \\ \text{thus} \quad & \widehat{F}_n^\star = F(\boldsymbol{P}^\star) \end{aligned}$$

- thus we have, $\widehat{\theta}_n^\star = t(\mathcal{X}_n^\star) = T(\widehat{F}_n^\star) = T(F(\boldsymbol{P}^\star)) = T(\boldsymbol{P}^\star)$

Bootstrap vs. Jackknife IV

- "jackknife is linearized bootstrap" in the following sense:

  Theorem 0.1. *Let us consider the problem of estimating $\theta = T(F)$ by the plug-in estimator $T(\widehat{F}_n) = t(\mathcal{X}_n) = \widehat{\theta}_n$. Here, $\widehat{\theta}_n$ may not a linear statistic. We know,*

$$Var_{jack}(t) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{\theta}_{(-i)} - \widehat{\theta}_{(\cdot)} \right)^2$$

  Let $T^{LIN}(\boldsymbol{P}^\star) := c_0 + (\boldsymbol{P}^\star - \boldsymbol{P}^0)^T \boldsymbol{U}$, where $\sum_{i=1}^n U_i = 0$ (so that only n among $c_0, \boldsymbol{U}$ are free to vary) be a linear statistic. So, $T^{LIN}(\boldsymbol{P}^\star)$ is a hyperplane defined on the n-dimensional simplex,
  $S_n := \{\boldsymbol{x} \in \mathbb{R}^n \mid 0 \le x_i \le 1,\ i = 1, 2, \ldots, n, \sum_{i=1}^n x_i = 1\}$. Solve for $c_0, \boldsymbol{U}$ under the n conditions $T(\boldsymbol{P}_{(i)}) = \widehat{\theta}_{(-i)},\ i = 1, 2, \ldots, n$ to get $T^{LIN, \widehat{\theta}_n}(\boldsymbol{P}^\star)$, the "linearized" version of $\widehat{\theta}^\star_n$. Then, for the ideal bootstrap variance of $T^{LIN, \widehat{\theta}_n}(\boldsymbol{P}^\star)$, we have

$$Var(T^{LIN, \widehat{\theta}_n}(\boldsymbol{P}^\star)) = \frac{n-1}{n} Var_{jack}(t)$$

Bootstrap vs. Jackknife V

- the set up (from problem 3 on problem set 5):
  - let $d > 1$ and $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)^T \sim \mathbf{Normal}_d(\boldsymbol{\mu}, \boldsymbol{I}_d)$, where $\mu_j = j,\ j = 1, 2, \ldots, d$ and $\boldsymbol{I}_d$ is the d-dimensional identity matrix
  - consider n independent copies $\boldsymbol{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})^T,\ i = 1, 2, \ldots, n$ of $\boldsymbol{X}$
  - let $\zeta_{1,j} := Var(\frac{1}{n}\sum_{i=1}^n X_{i,j})$ and $\zeta_{2,j} := Var\left(\frac{1}{n}\sum_{i=1}^n X_{i,j}^2\right)$ and define $\boldsymbol{\zeta}_k := (\zeta_{k,1}, \zeta_{k,2}, \ldots, \zeta_{k,d})^T,\ k = 1, 2.$
- note

$$
\begin{aligned}
Var_F(\widehat{\boldsymbol{\zeta}}_{boot,k,B,n}) &= Var_F(E_{\widehat{F}_n}(\widehat{\boldsymbol{\zeta}}_{boot,k,B,n} \mid \mathcal{X}_n)) + E_F(Var_{\widehat{F}_n}(\widehat{\boldsymbol{\zeta}}_{boot,k,B,n} \mid \mathcal{X}_n)) \\
&= Var_F(\widehat{\boldsymbol{\zeta}}_{k,n}) + E_F(Var_{\widehat{F}_n}(\widehat{\boldsymbol{\zeta}}_{boot,k,B,n} \mid \mathcal{X}_n)) \\
&\quad \text{if } E_{\widehat{F}_n}(\widehat{\boldsymbol{\zeta}}_{boot,k,B,n} \mid \mathcal{X}_n) = \widehat{\boldsymbol{\zeta}}_{k,n}
\end{aligned}
$$

Bootstrap vs. Jackknife VI

- thus bootstrap introduces the extra source of variation, called the *cushion error*, $E_F(Var_{\widehat{F}_n}(\widehat{\zeta}_{boot,k,B,n} \mid \mathcal{X}_n))$ due to finite re-sampling (of size B) from all possible bootstrap samples, see [4, Yatracos, 2002]
- the performance of $\widehat{\zeta}_{boot,k,B,n}$ will depend on the estimand (compare $\widehat{\zeta}_{boot,1,B,n}$ and $\widehat{\zeta}_{boot,2,B,n}$) and the dimension of the problem (e.g., compare $\widehat{\zeta}_{boot,1,B,n}$ for different values of the underlying d) since the both of these factors will contribute to the cushion error
- note for jackknife, the cushion error $E_F(Var_{\widehat{F}_n}(\widehat{\zeta}_{jack,k,n} \mid \mathcal{X}_n)) = 0$, since given a sample $\mathcal{X}_n$, there is only one jackknife estimator $\widehat{\zeta}_{jack,k,n}$!
- thus if we either use $\widehat{\zeta}_n$ or $\widehat{\zeta}_{jack,k,n}$ instead of $\widehat{\zeta}_{boot,k,B,n}$, then we'll not be paying for the cushion error

Bootstrap vs. Jackknife VII

- one way to bring the cushion error down is to use a very large value of B, especially for large dimensions, so that $Var_{\widehat{F}_n}(\widehat{\zeta}_{boot,k,B,n} \mid \mathcal{X}_n)$ is small
- another way to would be to only use biased set of bootstrap samples for which the $\widehat{F}_n^{\star}(\cdot)$'s lie "close" to the $\widehat{F}_n(\cdot)$ in some sense, see [3, Hall et. al., 1999]
- the later method is also called *biased-bootstrap* or *b-bootstrap*

References

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

[2] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall Ltd, 1993.

[3] Peter Hall and Brett Presnell. Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61:143–158, 1999.

[4] Yannis G. Yatracos. Assessing the quality of bootstrap samples and of the bootstrap estimates obtained with finite resampling. *Statistics & Probability Letters*, pages 281–292, 2002.

