Week 11: Bootstrap (continued)

MATH-517 Statistical Computation and Visualization

Tomas Masak

December 2nd 2022

The (standard/non-parametric) Bootstrap

- let $\mathcal{X} = \{X_1, \ldots, X_N\}$ be a random sample from F
- characteristic of interest: $\theta = \theta(F)$
- estimator: $\widehat{\theta} = \theta(\widehat{F}_N)$
  - write $\widehat{\theta} = \theta[\mathcal{X}]$, since $\widehat{F}_N$ and thus the estimator depend on the sample
- the distribution F_T of a scaled estimator $T = g(\widehat{\theta}, \theta) = g(\theta[\mathcal{X}], \theta)$ is of interest
  - e.g. $T = \sqrt{N}(\widehat{\theta} - \theta)$

The workflow of the bootstrap is as follows for some $B \in \mathbb{N}$ (e.g. $B = 1000$):

$$\begin{array}{ccl} \text{Data} & & \qquad\qquad\text{Resamples} \\ \mathcal{X} = \{X_1, \ldots, X_N\} & \Longrightarrow & \left\{ \begin{array}{ccc} \mathcal{X}_1^\star = \{X_{1,1}^\star, \ldots, X_{1,N}^\star\} & \Longrightarrow & T_1^\star = g(\theta[\mathcal{X}_1^\star], \theta[\mathcal{X}]) \\ \vdots & & \vdots \\ \mathcal{X}_B^\star = \{X_{B,1}^\star, \ldots, X_{B,N}^\star\} & \Longrightarrow & T_B^\star = g(\theta[\mathcal{X}_B^\star], \theta[\mathcal{X}]) \end{array} \right. \end{array}$$

F_T now estimated by $\widehat{F}^\star_{T,B}(x) = B^{-1} \sum_{b=1}^{B} \mathbb{I}_{[T_b^\star \leq x]}$

- any characteristic of F_T can be estimated by the char. of $\widehat{F}^\star_{T,B}(x)$

Confidence Intervals

- $T = \sqrt{N}(\widehat{\theta} - \theta)$ for $\theta \in \mathbb{R}$
- $T_b^\star = \sqrt{N}(\widehat{\theta}_b^\star - \widehat{\theta})$ for $b = 1, \ldots, B$

Asymptotic CI: $q(\alpha)$ is the α-quantile of the asymptotic distribution of T

$$\left(\widehat{\theta} - \frac{q(1-\alpha/2)}{\sqrt{N}}, \widehat{\theta} - \frac{q(\alpha/2)}{\sqrt{N}}\right)$$

Bootstrap CI: $q_B^\star(\alpha)$ is the empirical α-quantile of $\widehat{F}_{T,B}^\star$

$$\left(\widehat{\theta} - \frac{q_B^\star(1-\alpha/2)}{\sqrt{N}}, \widehat{\theta} - \frac{q_B^\star(\alpha/2)}{\sqrt{N}}\right)$$

Studentized CIs

Typically $\sqrt{N}(\widehat{\theta} - \theta) \to \mathcal{N}(0, v^2)$ for $\theta \in \mathbb{R}$

- let $\widehat{v}$ be a consistent estimator for v
- re-define $T = \sqrt{N}\frac{\widehat{\theta}-\theta}{\widehat{v}}$
  - $T_b^\star = \sqrt{N}\frac{\widehat{\theta}_b^\star - \widehat{\theta}}{\widehat{v_b^\star}}$ for $b = 1, \ldots, B$
  - this is called studentization, and is **always recommended** (sometimes provides better rates)
- asymptotic CI: $q(\alpha)$ is the α-quantile of $\mathcal{N}(0,1)$ (for the interval on the previous slide it would have been $\mathcal{N}(0, v^2)$)

$$\left(\widehat{\theta} - \frac{q(1-\alpha/2)}{\sqrt{N}}\widehat{v}, \widehat{\theta} - \frac{q(\alpha/2)}{\sqrt{N}}\widehat{v}\right)$$

- bootstrap CI: $q_B^\star(\alpha)$ is the empirical α-quantile of $\widehat{F}_{T,B}^\star$

$$\left(\widehat{\theta} - \frac{q_B^\star(1-\alpha/2)}{\sqrt{N}}\widehat{v}, \widehat{\theta} - \frac{q_B^\star(\alpha/2)}{\sqrt{N}}\widehat{v}\right)$$

Variance estimation

- often $\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}_p(0, \Sigma)$, but $\Sigma = \Sigma(\theta)$ complicated
- the bootstrap estimator of $N^{-1}\Sigma$ is easy to obtain:

$$\widehat{\Sigma}^\star = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}_b^\star - \bar{\theta}^\star\right)\left(\widehat{\theta}_b^\star - \bar{\theta}^\star\right)^\top, \qquad \text{where} \qquad \bar{\theta}^\star = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}_b^\star,$$

N^{-1} because one should take $T^\star = \sqrt{N}(\widehat{\theta}_b^\star - \widehat{\theta})$, and estimate Σ by

$$\frac{1}{B-1}\sum_{b=1}^{B}\left(T_b^\star - \bar{T}^\star\right)\left(T_b^\star - \bar{T}^\star\right)^\top \approx N^{-1}\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}_b^\star - \bar{\theta}^\star\right)\left(\widehat{\theta}_b^\star - \bar{\theta}^\star\right)^\top$$

Bias Reduction

- unbiased estimators are exception rather than a rule (apart from basic statistic classes)
- bootstrap estimates the bias as $\widehat{b}^\star = \bar{\theta}^\star - \widehat{\theta}$
- bias-corrected estimator defined as $\widehat{\theta}_b = \widehat{\theta} - \widehat{b}^\star$

Example: $X_1, \ldots, X_N$ are i.i.d. with $\mathbb{E}|X_1|^3 < \infty$, $\mathbb{E}X_1 = \mu$, and $\theta = \mu^3$. We saw last week

- $\widehat{\theta} = (\bar{X}_N)^3$
- $b := \text{bias}(\widehat{\theta}) = \mathbb{E}\widehat{\theta} - \theta$ is of order N^{-1}
- $\widehat{\theta}_b^\star = \widehat{\theta} - \widehat{b}^\star$ has bias of order N^{-2}

Something similar happens more generally for $\theta = g(\mu)$ when g is sufficiently smooth.

Hypothesis Testing

- testing H_0 using a statistic T
- depending on the form of the alternative H_1, evidence against H_0 is
  - large values of T,
  - small values of T, or
  - both large and small values of T
- bootstrap p-values
  - $\widehat{\text{p-val}} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}_{[T_b^\star \geq T]}\right)$,
  - $\widehat{\text{p-val}} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}_{[T_b^\star \leq T]}\right)$, or
  - $\widehat{\text{p-val}} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}_{[|T_b^\star| \geq |T|]}\right)$.

Example: $X_1, \ldots, X_N \overset{\perp\!\!\!\perp}{\sim} Exp(1/2)$ and $H_0 : \mu = 1.8$ is tested against

Example

the alternative $H_1 : \mu > 1.8$

```
set.seed(517)
N <- 100
X <- rexp(N,1/2)
mu_0 <- 1.8      # hypothesized value, so the hypothesis does not h
T_stat <- (mean(X)-mu_0)/sd(X)*sqrt(N)
B <- 10000
boot_stat <- rep(0,B)
for(b in 1:B){
  Xb <- sample(X,N,replace=T)
  boot_stat[b] <- (mean(Xb)-mean(X))/sd(Xb)*sqrt(N)
}
( p_val <- sum(boot_stat > T_stat)/(B+1) )
```

```
## [1] 0.04589541
```

```
1-pnorm(T_stat)
```

```
## [1] 0.0702328
```

Example

the alternative $H_1 : \mu \neq 1.8$

```
set.seed(517)
N <- 100
X <- rexp(N,1/2)
mu_0 <- 1.68     # reduced, since harder to reject here => hypothes
T_stat <- (mean(X)-mu_0)/sd(X)*sqrt(N)
B <- 10000
boot_stat <- rep(0,B)
for(b in 1:B){
  Xb <- sample(X,N,replace=T)
  boot_stat[b] <- (mean(Xb)-mean(X))/sd(Xb)*sqrt(N)
}
( p_val <- sum(abs(boot_stat) > abs(T_stat))/(B+1) )
```

```
## [1] 0.06129387
```

```
2*(1-pnorm(T_stat))
```

```
## [1] 0.0455959
```

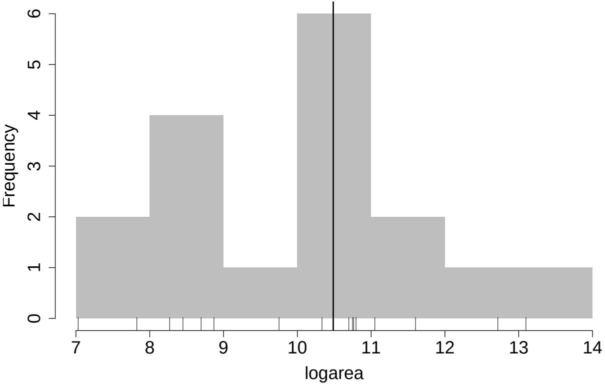
Example: Median

- Antarctic ice shelves data
- interested in the median of the log-area of the ice shelves

```
aa <- read.csv('../data/AAshelves.csv')
  # source: Reinhard Furrer's "Statistical Modeling" lecture at UZH
logarea <- log(aa[[3]]) # log of ice shelf areas
set.seed(517)
N <- 17
B <- 5000
boot_data <- array(sample(logarea, N*B, replace=TRUE), c(B, N))
meds <- apply(boot_data, 1, median)
hist(logarea, col='gray', main='', border=NA)
rug(logarea, ticksize = .04)
abline(v=median(logarea),lwd=2)
```

Example: Median

- Antarctic ice shelves data
- interested in the median of the log-area of the ice shelves



Example: Median

Is the sample median asymptotically normal?

$$\theta = F^{-1}(1/2) \qquad \& \qquad \widehat{\theta} = \widehat{F}_N^{-1}(1/2)$$

$T = \sqrt{N}(\widehat{\theta} - \theta) \stackrel{?}{\to} \mathcal{N}(0, v)$

Example: Median

Is the sample median asymptotically normal?

$$\theta = F^{-1}(1/2) \qquad \& \qquad \widehat{\theta} = \widehat{F}_N^{-1}(1/2)$$

$T = \sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{?} \mathcal{N}(0, v)$

- yes, under some conditions
  - verifying conditions of a general theorem for M-estimator yields assumption:
  - $f(\theta) \neq 0$ and f continuous on some neighborhood of θ

Say we wish to construct a confidence interval.

Option I:

- approximate only v using bootstrap

Option II:

- approximate the quantiles of T using bootstrap

Example: Median

$T^\star = \sqrt{N}(\widehat{\theta}^\star - \widehat{\theta})$ or just $T^\star = \widehat{\theta}^\star$

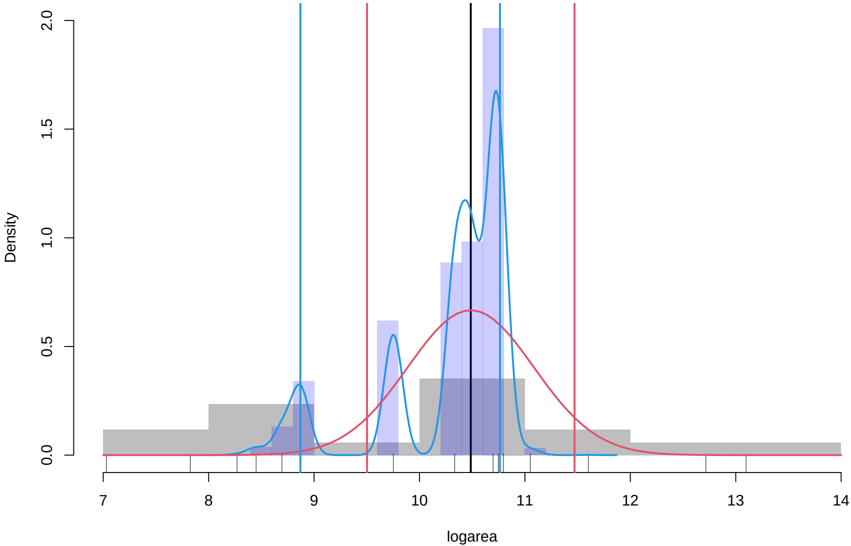
Option I: approximate $\mathrm{avar}(T^\star)$ using MC

Option II: approximate the quantiles of $T^\star$ using MC

- KDE on the MC draws of $T^\star$ can be used to visualize the distribution

```
hist(logarea, prob=TRUE, col='gray', ylim=c(0,2.), main='', border=NA)
rug(logarea, ticksize = .04)
abline(v=median(logarea),lwd=2)
hist(meds, add=T, prob=T, col=rgb(0,0,1,.2), border=NA)
lines(density(meds, adjust=2), col=4, lwd=2)
curve(dnorm(x, median(logarea), sd(meds)), add=T, col=2,lwd=2)
abline(v=c(median(logarea)+qnorm(c(.05,.95))*sd(meds)), col=2, lwd=2) # I
# sd(meds) == sd(sqrt(N)*(meds-median(logarea)))/sqrt(N)
abline(v=c(quantile(meds, c(.05,.95))), col=4, lwd=2) ## II
```

Example: Median



Iterated Bootstrap

Simple bootstrap:

$$
\begin{array}{ccccccc}
\text{Data} & & \text{Resamples} & & & & \\
\mathcal{X} = \{X_1, \ldots, X_N\} & \Longrightarrow & \left\{ \begin{array}{l} \mathcal{X}_1^\star = \{X_{1,1}^\star, \ldots, X_{1,N}^\star\} \quad \Longrightarrow \quad T_1^\star = g(\theta[\mathcal{X}_1^\star], \theta[\mathcal{X}]) \\ \qquad \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ \mathcal{X}_B^\star = \{X_{B,1}^\star, \ldots, X_{B,N}^\star\} \quad \Longrightarrow \quad T_B^\star = g(\theta[\mathcal{X}_B^\star], \theta[\mathcal{X}]) \end{array} \right. & & & &
\end{array}
$$

Iterated Bootstrap

Double bootstrap:

$$
\begin{array}{ccccc}
\text{Data} & & \text{Resamples} & & \text{Re-resamples} \\
\mathcal{X} & \Longrightarrow & \left\{ \begin{array}{c} \mathcal{X}_1^\star \quad \Longrightarrow \left\{ \begin{array}{c} \mathcal{X}_{1,1}^{\star\star} \\ \vdots \\ \mathcal{X}_{1,C}^{\star\star} \end{array} \right. \\ \\ \vdots \qquad\qquad\qquad \vdots \\ \\ \mathcal{X}_B^\star \quad \Longrightarrow \left\{ \begin{array}{c} \mathcal{X}_{B,1}^{\star\star} \\ \vdots \\ \mathcal{X}_{B,C}^{\star\star} \end{array} \right. \end{array} \right. & &
\end{array}
$$

Example

- $X_1, \ldots, X_p \in \mathbb{R}^p$ be i.i.d. from a distribution depending on $\theta \in \mathbb{R}^p$
- $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$
- $\widehat{\theta}$ satisfies $\sqrt{N}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$
- studentized statistic:

$$T = \sqrt{N}\widehat{\Sigma}^{-1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_{p \times p}) \qquad \text{(under } H_0\text{)}$$

  - $\widehat{\Sigma}$ is consistent for Σ
- asymptotic test based on: $\|T\|^2 \xrightarrow{d} \chi^2_p$ under H_0

Bootstrap can be used

- instead of using the asymptotic distribution to produce a p-value, or
- when an estimator of Σ is not available

Both of the above combined $\Rightarrow$ double bootstrap

Example

$$
\mathcal{X} = \{X_1,\ldots,X_N\}
\left\{
\begin{array}{lcccc}
\mathcal{X}_1^\star = \{X_{1,1}^\star,\ldots,X_{1,N}^\star\} & \left\{\begin{array}{c} \mathcal{X}_{1,1}^{\star\star} = \{X_{1,1,1}^{\star\star},\ldots,X_{1,1,N}^{\star\star}\} \\ \vdots \\ \mathcal{X}_{1,M}^{\star\star} = \{X_{1,M,1}^{\star\star},\ldots,X_{1,M,N}^{\star\star}\} \end{array}\right\} & \widehat{\Sigma}_1^{\star\star} & \Longrightarrow & T_1^\star \\
 & \vdots & & & \\
\vdots & \vdots & & & \\
 & \vdots & & & \\
\mathcal{X}_B^\star = \{X_{B,1}^\star,\ldots,X_{B,N}^\star\} & \left\{\begin{array}{c} \mathcal{X}_{B,1}^{\star\star} = \{X_{B,1,1}^{\star\star},\ldots,X_{B,1,N}^{\star\star}\} \\ \vdots \\ \mathcal{X}_{B,M}^{\star\star} = \{X_{B,M,1}^{\star\star},\ldots,X_{B,M,N}^{\star\star}\} \end{array}\right\} & \widehat{\Sigma}_B^{\star\star} & \Longrightarrow & T_B^\star
\end{array}
\right\} \widehat{\text{p-val}}
$$

$$
\widehat{\Sigma}_b^{\star\star} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_{b,m}^{\star\star} - \bar{\theta}_b^{\star\star}\right)\left(\hat{\theta}_{b,m}^{\star\star} - \bar{\theta}_b^{\star\star}\right)^\top, \quad \text{where} \quad \hat{\theta}_m^{\star\star} = \theta\big[\mathcal{X}_{b,m}^{\star\star}\big] \quad \& \quad \bar{\theta}_b^{\star\star} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{b,m}^{\star\star},
$$

$$
T_b^\star = \sqrt{N}\left(\widehat{\Sigma}_b^{\star\star}\right)^{-1/2}\left(\hat{\theta}_b^\star - \hat{\theta}\right),
$$

$$
\widehat{\text{p-val}} = \frac{1}{1+B}\left(1 + \sum_{b=1}^{B} I\big(\|T_b^\star\|^2 \geq \|T\|^2\big)\right),
$$

Example: Median (continued)

Goal: construct CI for the median

Option I: approximate only the asymptotic variance v using bootstrap

- asymptotic

Option II: approximate directly the quantiles of using bootstrap

- non-studentized CI

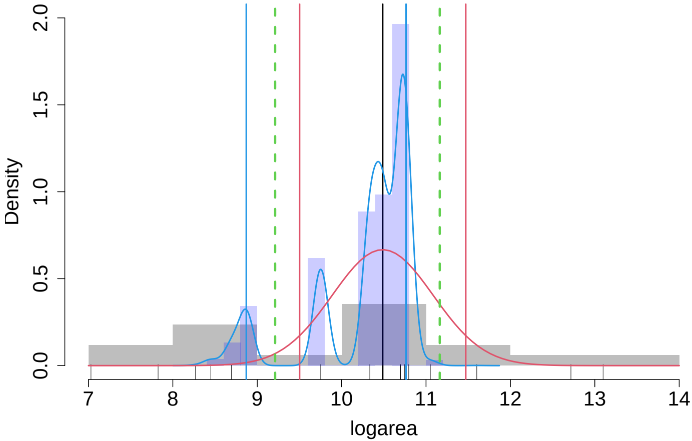
Option III: approximate the quantiles of a studentized statistic using one bootstrap (requires the knowledge of variance, so get that by using another bootstrap)

- studentized CI

Example: Median (continued)

```
set.seed(517)
N <- 17; B <- 5000; C <- 500;
boot_data <- array(sample(logarea, N*B, replace=TRUE), c(B, N))
# Dboot_data <- array(0,c(B,C,N))
# for(b in 1:B){
#   Dboot_data[b,,] <- array(sample(boot_data[b,], N*C, replace=TRUE), c(C, N))
# }
# meds <- apply(boot_data, 1, median)
# Dmeds <- apply(Dboot_data, c(1,2), median)
# sds <- apply(Dmeds, 1, sd)
# T_stars <- sqrt(N)*(meds - median(logarea))/sds
op <- par(ps=20)
hist(logarea, prob=TRUE, col='gray', ylim=c(0,2.), main='', border=NA)
rug(logarea, ticksize = .04)
abline(v=median(logarea),lwd=2)
hist(meds, add=T, prob=T, col=rgb(0,0,1,.2), border=NA)
lines(density(meds, adjust=2), col=4, lwd=2)
curve(dnorm(x, median(logarea), sd(meds)), add=T, col=2,lwd=2)
abline(v=median(logarea)+qnorm(c(.05,.95))*sd(meds), col=2, lwd=2)
### sd(meds) == sd(sqrt(N)*(meds-median(logarea)))/sqrt(N)
abline(v=quantile(meds, c(.05,.95)), col=4, lwd=2)
# abline(v=median(logarea)+quantile(T_stars, c(.05,.95))/sqrt(N)*sd(meds), col=3, lw
abline(v=c(9.212299, 11.162336), col=3, lwd=2, lty=2) # studentized CI
```

Example: Median (continued)



Is the studentized CI actually better? Simulations!

Parametric Bootstrap and GoF Testing

- $X_1, \ldots, X_N \overset{\perp\!\!\!\perp}{\sim} F$
- **goal**: test $H_0 : F \in \mathcal{F} = \{F_\lambda \mid \lambda \in \Lambda\}$ against $H_1 : F \notin \mathcal{F}$
  - if $\mathcal{F} = \{F_0\}$, we could use the KS statistic: $\sup_x \left| \widehat{F}_N(x) - F_0(x) \right|$
- plug in principle: use $T = \sup_x \left| \widehat{F}_N(x) - F_{\widehat{\lambda}}(x) \right|$
  - where $\widehat{\lambda}$ is consistent under H_0 (e.g. the MLE)

Bootstrap procedure: **for** $b = 1, \ldots, B$

- generate $\mathcal{X}_b^\star = \{X_{b,1}^\star, \ldots, X_{b,N}^\star\}$
  - this time not by resampling, but by sampling from $F_{\widehat{\lambda}}$
- estimate $\widehat{\lambda}_b^\star$ from $\mathcal{X}_b^\star$
- calculate the EDF $\widehat{F}_{N,b}^\star$ from $\mathcal{X}_b^\star$
- set $T_b^\star = \sup_x \left| \widehat{F}_{N,b}^\star(x) - F_{\widehat{\lambda}_b^\star}(x) \right|$

Jackknife

- a predecessor to the bootstrap
  - sometimes can achieve a better trade-off between accuracy and computational costs, but hard to quantify
- used first for bias correction, later for variance estimation

$X_1, \ldots, X_N$ a random sample from F depending on $\theta \in \mathbb{R}^p$

- $\widehat{\theta} = \theta[X_1, \ldots, X_N]$
  - interested in some characteristic of the estimator such as the bias
- consider $\bar{\theta} = N^{-1} \sum_n \widehat{\theta}_{-n}$, where $\widehat{\theta}_{-n} = \theta[X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_N]$

Jackknife estimator of the bias:

$$\widehat{b} = (N-1)(\bar{\theta} - \widehat{\theta})$$

- the scaling factor suprising?

Jackknife Bias - a Heuristic

- assume $b = \text{bias}(\widehat{\theta}) = aN^{-1} + bN^{-2} + \mathcal{O}(N^{-3})$ for some constants a and b

$$\text{bias}(\widehat{\theta}_{-n}) = a(N-1)^{-1} + b(N-1)^{-2} + \mathcal{O}(N^{-3}) = \text{bias}(\bar{\theta}).$$

$$\begin{aligned}
\mathbb{E}\widehat{b} &= (N-1)\big[\text{bias}(\bar{\theta}) - \text{bias}(\widehat{\theta})\big] \\
&= (N-1)\left[a\left(\frac{1}{N-1} - \frac{1}{N}\right) + \left(\frac{1}{(N-1)^2} - \frac{1}{N^2}\right) + \mathcal{O}\left(\frac{1}{N^3}\right)\right], \\
&= aN^{-1} + bN^{-2}\frac{2N-1}{N-1} + \mathcal{O}(N^{-3})
\end{aligned}$$

- so $\widehat{b}$ approximates b correctly up to the order N^{-1}, which corresponds to the bootstrap
    - and similarly $\widehat{\theta}_b^\star = \widehat{\theta} - \widehat{b} = N\widehat{\theta} - (N-1)\bar{\theta}$ has bias of order N^{-1}, etc.

Jackknife Variance

John W. Tukey defined the "pseudo-values"

$$\theta_n^\star = N\widehat{\theta} - (N-1)\widehat{\theta}_{-n}$$

and conjectured that in some situations these can be treated as i.i.d. with approximately the same variance as $N\mathrm{var}(\widehat{\theta})$, and hence we can take

$$\widehat{\mathrm{var}}(\widehat{\theta}) = \frac{1}{N}\frac{1}{N-1}\sum_{n=1}^{N}\left(\theta_n^\star - \bar{\theta}^\star\right)\left(\theta_n^\star - \bar{\theta}^\star\right)^\top.$$

- later shown to actually work (studying the theoretical version of the jackknife)
- could be used instead of the second bootstrap in our double bootstrap example above

Assignment 7 [5 %]

For $X_1, \ldots, X_{100} \overset{\perp\!\!\!\perp}{\sim} Exp(2)$, consider the following CIs for $\mathbb{E}X_1 = 1/2$:

- asymptotic: $\left(-\infty, \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} z(\alpha)\right)$
- studentized (bootstrap): $\left(-\infty, \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} q^\star(\alpha)\right)$ with $T^\star = \sqrt{N}\frac{\bar{X}_N^\star - \bar{X}_N}{\widehat{\sigma}^\star}$
- non-studentized ($\star$): $\left(-\infty, \bar{X}_N + \frac{1}{\sqrt{N}} q^\star(\alpha)\right)$ with $T^\star = \sqrt{N}\big(\bar{X}_N^\star - \bar{X}_N\big)$
- sample-truth-scaled ($\star$): $\left(-\infty, \bar{X}_N + \frac{\widehat{\sigma}}{\sqrt{N}} q^\star(\alpha)\right)$ with $T^\star = \sqrt{N}\frac{\bar{X}_N^\star - \bar{X}_N}{\widehat{\sigma}}$

Verify coverage of these intervals via a simulation study of 10^3 runs and report the coverage proportions as a table. Specifically, for every single one of 10^3 simulation runs:

- generate new data $X_1, \ldots, X_{100} \overset{\perp\!\!\!\perp}{\sim} Exp(2)$
- calculate the four confidence intervals
- check whether $\mathbb{E}X_1 = 1/2$ lies inside the respective intervals (yes = coverage)
- report the coverage proportion for the respective intervals as a single table