

450A Political Methodology I

Lecture 2: Univariate Statistical Inference

4. Additional Topics

Stanford University

Motivation

Bootstrap

Clusters

Optional: Estimating $\text{Var}(\bar{X})$ in Practice

Optional: Better Bootstrap Intervals

Uncertainty Estimates for Quantities of Interest

Consider a general inference problem:

- An i.i.d. sample of size n : $X = \{X_1, \dots, X_n\}$
- An unknown population distribution P : $X_i \stackrel{\text{i.i.d.}}{\sim} P$
- Let $\hat{\theta} = s(X)$ be a statistic that estimates a parameter θ of the population distribution P
- Without assuming a parametric form for P , how can we make inferences about the sampling distribution of our statistic $\hat{\theta}$ to determine its bias and variance or to construct confidence intervals?

Uncertainty Estimates for Sample Mean Estimator

Assume our statistic of interest is the sample mean:

$$\hat{\theta} = s(X) = \frac{1}{n} \sum_i X_i = \bar{X}_n$$

- assume that P has mean μ and variance σ^2
- then we know from the CLT that

$$\hat{\theta} \stackrel{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- which allows us to estimate standard errors and confidence intervals using the usual formulas. For standard errors we use

$$\hat{SE}[\hat{\theta}] = \frac{S}{\sqrt{n}} \text{ where } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

and for confidence intervals

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

Now instead assume our statistic of interest is the sample median:

$$\hat{\theta} = s(X) = \text{med}(X)$$

- then even if we assume that P has a median m_P it is not straightforward to derive the sampling distribution of $\hat{\theta}$
- similar issues apply if our statistic of interest is a more complex quantity such as a percentile, a ratio, or a correlation coefficient

Motivation

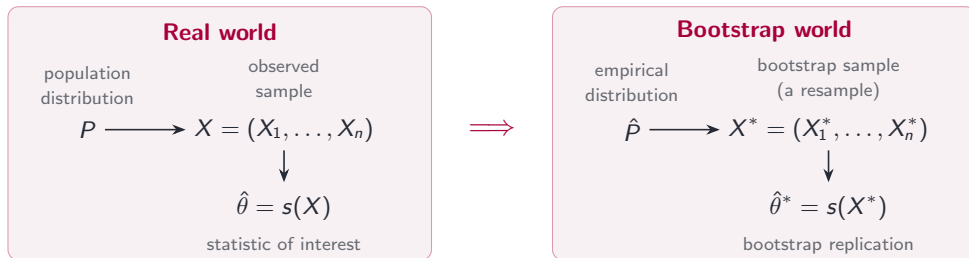
Bootstrap

Clusters

Optional: Estimating $\text{Var}(\bar{X})$ in Practice

Optional: Better Bootstrap Intervals

Welcome to the Bootstrap World



- Given no other information, the observed sample X contains all the available information about the underlying population distribution P , and so **resampling** from X is the best guide to what can be expected from resampling from P
- Just like we drew X from P , let's draw a **resample** X^* from X

Source: Diagram after Efron & Tibshirani (1993)

Welcome to the Bootstrap World

- If n is sufficiently large, the sample X should be a good approximation of P
- In fact, the empirical distribution of the sample X , call it $\hat{P}$, is a non-parametric maximum likelihood estimator of the population P .
- $\hat{P}$ is a discrete probability distribution that puts mass $1/n$ on each observation, so a value observed k times gets mass k/n
- By the LLN, $\hat{P}(A) \xrightarrow{P} P(A)$ for each set A as $n \rightarrow \infty$

Nonparametric Bootstrap

You can do this without assuming a parametric form for P , and for most smooth functions of the data

$$\hat{\theta} = s(X)$$

1. Draw B resamples of size n from X with replacement
2. For each X_b^* , compute $\hat{\theta}_b^* = s(X_b^*)$
- 3a. To obtain s.e. of $\hat{\theta}$, use the sample standard deviation of the replications $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ (**bootstrap standard errors**)
- 3b. To compute 95% CI, use 2.5/97.5 percentiles of the replications $\{\hat{\theta}_1^*, \dots, \hat{\theta}_B^*\}$ as the lower/upper bounds (**bootstrap percentile CI**)
- 3c. If you know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}$, you can use 3a. and compute the **bootstrap normal CI**
- 3d. Optionally, *Bootstrap-t*: $t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\widehat{\text{se}}(\hat{\theta}_b^*)}$, which is approximately “pivotal”; take percentiles of t_b^* (optional slides at the end).

Parametric Bootstrap

Assume $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\psi$, a parametric family with density $f(x; \psi)$, and a statistic of interest $\hat{\theta} = s(X)$.

1. Estimate the model parameter ψ (e.g., by MLE) and compute $\hat{\theta} = s(X)$.
2. For $b = 1, \dots, B$:
 - (a) Draw a synthetic sample $X_b^* = (X_{b1}^*, \dots, X_{bn}^*) \stackrel{\text{i.i.d.}}{\sim} P_{\hat{\psi}}$.

For example, in a regression setting: if $Y_i = W_i\beta + \varepsilon_i$, simulate $Y_i^* = W_i\hat{\beta} + \varepsilon_i^*$ with $\varepsilon_i^* \sim F_{\varepsilon; \hat{\eta}}$ (η : the parameters of the error distribution), e.g., $\mathcal{N}(0, \hat{\sigma}^2)$, then re-estimate; this is “adding random noise” under the assumed error model.

- (b) Compute $\hat{\theta}_b^* = s(X_b^*)$.
3. Inference using bootstrapped SE, CI, or t statistics

→ Often more efficient than nonparametric bootstrap when P_ψ is well specified; but sensitive to misspecification.

Additional Resources:

- ① Bradley Efron and R.J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1993).
- ② A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application* (Cambridge Series in Statistical and Probabilistic Mathematics, 1997).
- ③ Peter Hall, *The Bootstrap and Edgeworth Expansion* (Springer Series in Statistics, Springer, 1992).

Motivation

Bootstrap

Clusters

Optional: Estimating $\text{Var}(\bar{X})$ in Practice

Optional: Better Bootstrap Intervals

Clustered Data Structure

Consider a common characteristic in many datasets:

- A sample of size n : $X = \{X_1, \dots, X_n\}$
- But the sample is not completely i.i.d.
- Instead, the sample contains m_g individuals from G separate clusters.
- We assume that the set of G clusters included in the sample were randomly sampled from $\mathcal{G}$ clusters in the population.
- The problem is that there is often more homogeneity among individuals within cluster than across clusters, such that individuals within a cluster cannot be treated as independently sampled observations.
- The result: when the within-cluster correlation is positive, inference based on i.i.d. sampling assumptions produces standard errors that are biased downward.

Worksheet Q1

Suppose the sample consists of G independent clusters, each of size m (so $n = mG$). Consider the one-way random-effects model

$$X_{gi} = \mu + u_g + \varepsilon_{gi}, \quad u_g \stackrel{iid}{\sim} (0, \sigma_B^2), \quad \varepsilon_{gi} \stackrel{iid}{\sim} (0, \sigma_W^2),$$

with $u_g \perp \varepsilon_{gi}$. Let $\sigma^2 = \sigma_B^2 + \sigma_W^2$ and the **intra-cluster correlation** $\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$.

Show that

$$\text{Var}(\bar{X}) = \frac{\sigma^2 + (m-1)\sigma_B^2}{mG} = \frac{\sigma^2}{mG} (1 + (m-1)\rho).$$

- With clustered data, the population variance σ^2 can be decomposed into:
 - Variance between clusters σ_B^2 (i.e. variance of the clusters' population means $\mu + u_g$)
 - Variance within cluster σ_W^2 (i.e. individual variance within cluster)
 - If clusters have the same mean: $\rho = 0 \rightarrow$ regular i.i.d. sampling assumptions suffice
- As a result of intra-cluster correlation, $\text{Var}(\bar{X}) \neq \frac{\sigma^2}{n}$.
- Instead, assuming equal sized clusters (i.e. $m_g = m$):

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{mG} \left(1 + (m-1)\rho \right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n}(m-1)\rho$$

- With unequal cluster sizes, formulas become messier but the ideas and intuition are the same.
- How to estimate ρ and $\text{Var}(\bar{X})$ from data (ANOVA mean squares, Worksheet Q2): see the optional slides at the end.

Example: Gerber Voter Turnout Data

Did registered voters turn out in the August 2006 Michigan primary? Data from the Gerber, Green & Larimer (2008) mailing experiment.

- Course subsample: $n = 14,392$ voters in $G = 7,599$ households, every voter in the study data from each sampled household included; average household size $m = 1.89$
- The clusters are households: members of the same household tend to vote, or stay home, together
- Turnout $\bar{X} = 0.3232$; a one-way ANOVA with households as the groups gives $\hat{\rho} = 0.50$ (the estimator is in the optional slides at the end)

	Standard error	95% confidence interval
Conventional (i.i.d.)	0.00390	[0.3155, 0.3308]
Cluster-adjusted (average household size)	0.00470	[0.3140, 0.3324]

Treating the 14,392 voters as independent understates the standard error by 17% relative to this average-size adjustment, and by 19% relative to the cluster-robust SE with actual household sizes (0.00482).

Source: Gerber, Green & Larimer (2008), *American Political Science Review* 102(1)

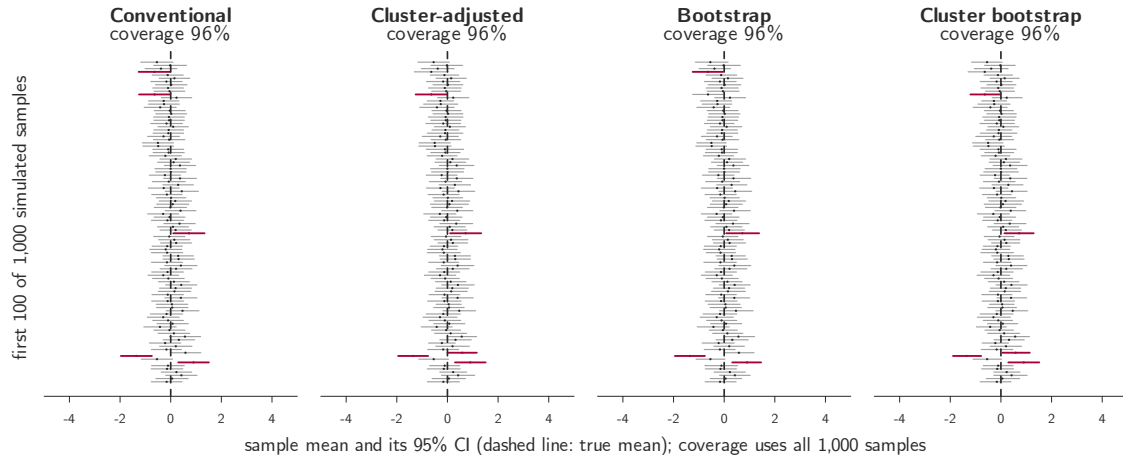
Alternative Estimation: Cluster/Block Bootstrap

- Rather than relying upon an analytical formula, we can instead employ bootstrap methods.
- However, for proper inference, it is vital that the bootstrap procedure mirrors the clustered sampling process / data structure.
- That is, instead of resampling individuals (with replacement), we **resample entire clusters (with replacement)**:
 0. Call $\mathbf{X}$ your entire sample of n observations. Columns of $\mathbf{X}$ constitute different clusters. For balanced clusters, $\mathbf{X}$ is a matrix with m rows and G columns.
 1. Draw B resamples of size G from the columns of $\mathbf{X}$, with replacement (i.e. draw entire columns). Hence, each of your B resamples will contain $mG = n$ observations (the same size as the sample).
 2. Proceed with bootstrap inference as before.

Note: This procedure can be adapted to unequal cluster sizes.

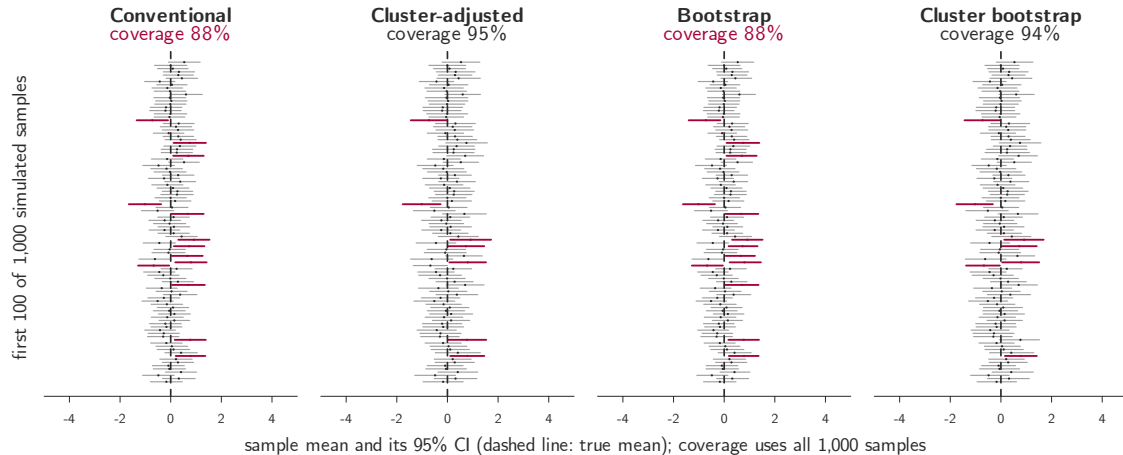
Relative Performance of Different Inference Procedures

Given normally distributed data, $\rho = 0$, $m = 10$, $G = 100$, $\alpha = 0.05$



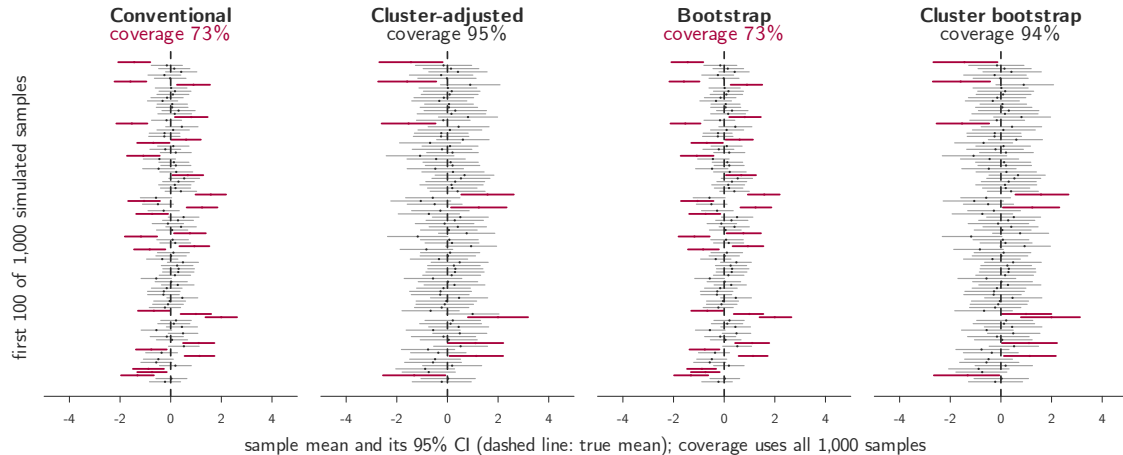
Relative Performance of Different Inference Procedures

Given normally distributed clustered data, $\rho = 0.05$, $m = 10$, $G = 100$, $\alpha = 0.05$



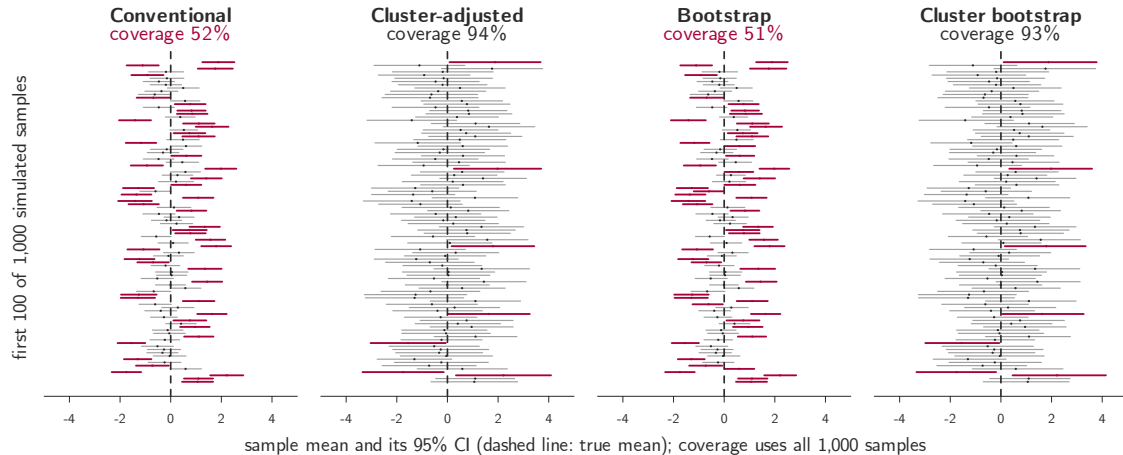
Relative Performance of Different Inference Procedures

Given normally distributed clustered data, $\rho = 0.25$, $m = 10$, $G = 100$, $\alpha = 0.05$



Relative Performance of Different Inference Procedures

Given normally distributed clustered data, $\rho = 0.75$, $m = 10$, $G = 100$, $\alpha = 0.05$



Motivation

Bootstrap

Clusters

Optional: Estimating $\text{Var}(\bar{X})$ in Practice

Optional: Better Bootstrap Intervals

(Optional) Estimating $\text{Var}(\bar{X})$ in Practice

Under the same setup with equal cluster sizes, define the cluster means $\bar{X}_g = \frac{1}{m} \sum_{i=1}^m X_{gi}$ and the grand mean $\bar{X} = \frac{1}{G} \sum_{g=1}^G \bar{X}_g$. Let

$$\text{MSW} = \frac{1}{G(m-1)} \sum_{g=1}^G \sum_{i=1}^m (X_{gi} - \bar{X}_g)^2, \quad \text{MSB} = \frac{m}{G-1} \sum_{g=1}^G (\bar{X}_g - \bar{X})^2.$$

Standard (though rather complicated) ANOVA shows that $\mathbb{E}[\text{MSW}] = \sigma_W^2$ and $\mathbb{E}[\text{MSB}] = \sigma_W^2 + m\sigma_B^2$.

Worksheet Q2

- (a) Propose consistent estimators for σ_W^2 , σ_B^2 , and the intraclass correlation coefficient (ICC)
 $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_W^2)$.
- (b) Using your results from (a), derive an estimator $\widehat{\text{Var}}(\bar{X})$ for the sampling variance of the sample mean $\bar{X}$.
- (c) Show that $\hat{\rho} \xrightarrow{P} \rho$ and $\widehat{\text{Var}}(\bar{X}) / \text{Var}(\bar{X}) \xrightarrow{P} 1$ as $G \rightarrow \infty$ with fixed m .
- (d) Construct a $100(1 - \alpha)\%$ cluster-adjusted confidence interval for μ .

(Optional) Estimating $\text{Var}(\bar{X})$ in Practice

- Because $\{\bar{X}_g\}_{g=1}^G$ behave like G approximately independent draws with variance $\sigma_B^2 + \sigma_W^2/m$, small-sample inference for μ should use $(G - 1)$ degrees of freedom. A cluster-adjusted CI is therefore

$$\left[\bar{X} \pm t_{\alpha/2, (G-1)} \sqrt{\widehat{\text{Var}}(\bar{X})} \right].$$

- Sanity checks:
 - If $\text{MSB} = \text{MSW}$ then $\hat{\rho} = 0$ and $\widehat{\text{Var}}(\bar{X}) = \hat{\sigma}^2/n$ with $\hat{\sigma}^2 = \hat{\sigma}_B^2 + \hat{\sigma}_W^2$ (the i.i.d. formula).
 - If $\text{MSB} > \text{MSW}$, then $\hat{\rho} > 0$, the design effect $1 + (m - 1)\hat{\rho} > 1$, so this CI is (appropriately) wider than the conventional i.i.d. CI.
- In R, use the `anova()` function to obtain mean squares estimates. For better understanding, review ANOVA in greater depth.
- Again, unbalanced cluster sizes make formulas messier. In practice, people often use the formulas above and set m equal to the *average sampled cluster size*. The more equal are the sampled cluster sizes, the more reasonable will be this simplification.

Motivation

Bootstrap

Clusters

Optional: Estimating $\text{Var}(\bar{X})$ in Practice

Optional: Better Bootstrap Intervals

(Optional) Percentile Interval: Recenter and Reverse

The percentile CI treats the replications $\hat{\theta}_b^*$ as if they were draws of $\hat{\theta}$ around θ . The bootstrap analogy is about *errors*: $\hat{\theta}^* - \hat{\theta}$ (resample vs. sample) mimics $\hat{\theta} - \theta$ (sample vs. population).

- **Recenter.** With $q_{.025}^*, q_{.975}^*$ the percentiles of the $\hat{\theta}_b^*$:

$$\Pr^* (q_{.025}^* - \hat{\theta} \leq \hat{\theta}^* - \hat{\theta} \leq q_{.975}^* - \hat{\theta}) = 0.95 \implies \Pr (q_{.025}^* - \hat{\theta} \leq \hat{\theta} - \theta \leq q_{.975}^* - \hat{\theta}) \approx 0.95$$

- **Reverse.** Solve for θ : $[2\hat{\theta} - q_{.975}^*, 2\hat{\theta} - q_{.025}^*]$ (the **basic**, or reversed percentile, CI)
 - A long right tail in the $\hat{\theta}_b^*$ becomes a long *left* arm of the interval
 - Bias in the $\hat{\theta}_b^*$ is subtracted rather than added
 - If the $\hat{\theta}_b^*$ are symmetric around $\hat{\theta}$, the two intervals coincide
- Median of the 20 incomes from Lecture 2-1 ($\hat{\theta} = 65.25$): percentile CI [43.5, 97.4]; reversed CI [33.1, 87.0]
- Which one is right depends on the statistic
 - percentile CI: right if some monotone transformation h makes $h(\hat{\theta}) - h(\theta)$ symmetric around 0, with the same distribution whatever θ is
 - reversed CI: right if $\hat{\theta} - \theta$ has the same distribution whatever θ is, e.g., a statistic with a stable bias

(Optional) When Reversal Helps: A Biased Statistic

The **Gini coefficient** of 20 incomes drawn from the Lecture 2-1 population (true Gini 0.273)

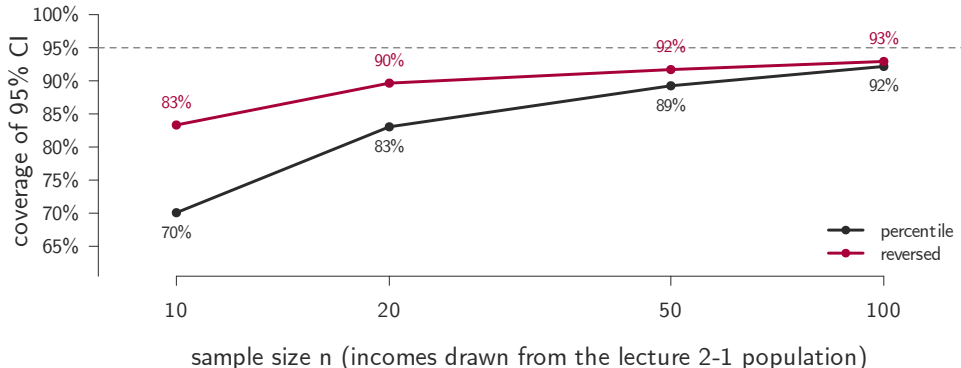
Reversed CI [0.250, 0.396]: flipping each endpoint around $\widehat{\text{Gini}}$ moves the interval up and **removes** the bias.



(Optional) When Reversal Helps: A Biased Statistic

The **Gini coefficient** of 20 incomes drawn from the Lecture 2-1 population (true Gini 0.273)

Coverage over 4,000 samples per n : the reversed CI misses less, most of all in small samples. The gain is not universal: for the median ($n = 20$), percentile 93%, reversed 81%; for a correlation of 0.8, percentile 92%, reversed 85%.



(Optional) Refinement: Bootstrap- t

The reversed CI assumes $\hat{\theta} - \theta$ has the same *spread* in every sample. Often it does not: with right-skewed incomes, samples with a high mean also have a large S .

- **Studentize** each replication, using an SE computed inside resample b (here $S_b^*/\sqrt{n}$):

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\widehat{se}_b} \quad \text{mimics} \quad \frac{\hat{\theta} - \theta}{\widehat{se}}$$

This ratio is approximately **pivotal**: its distribution hardly depends on unknowns

- Recenter and reverse as before, with the percentiles of the t_b^* in place of the t -table:

$$[\hat{\theta} - t_{.975}^* \widehat{se}, \hat{\theta} - t_{.025}^* \widehat{se}]$$

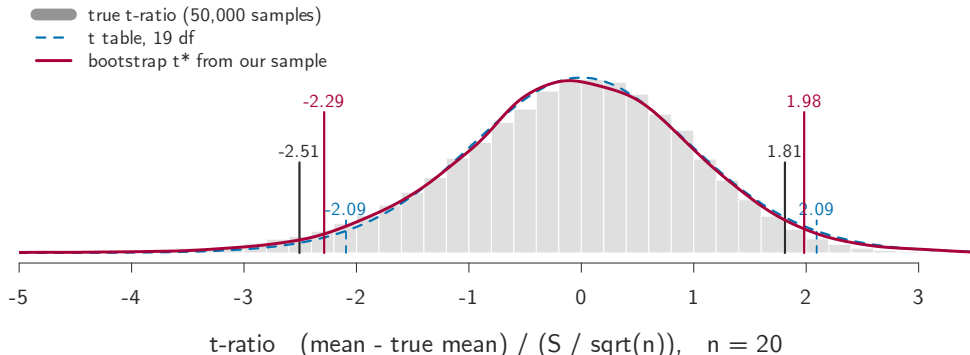
- Mean of the 20 incomes: $\hat{\theta} = 72.76$, $\widehat{se} = 9.04$; $t_{.025}^* = -2.29$, $t_{.975}^* = 1.98$ (the t_{19} table: ± 2.09)
 - bootstrap- t CI [54.8, 93.4], longer on the right; t CI [53.8, 91.7]; percentile [55.7, 90.4]
- Why it helps: under regularity conditions, each tail's coverage error shrinks like $1/n$, not $1/\sqrt{n}$ as for the percentile and reversed CIs (Hall 1992)
- Cost: an SE in every resample (for a median, a bootstrap inside each bootstrap). Alternative: BCa, `boot::boot.ci(type = "bca")`

Source: Hall (1992), *The Bootstrap and Edgeworth Expansion*, Springer, Ch. 3

(Optional) When Bootstrap- t Helps: Skewed Data

The mean of right-skewed incomes: samples of 20 from the Lecture 2-1 population (true mean 72.0)

The bootstrap t^* from our one sample **catches much of the skew**: -2.29 and 1.98 .



(Optional) When Bootstrap- t Helps: Skewed Data

The mean of right-skewed incomes: samples of 20 from the Lecture 2-1 population (true mean 72.0)

Coverage over 10,000 samples per n . At $n = 20$: bootstrap- t 94.7%, t 94.1%, percentile 91.8%, reversed 91.3%. Bootstrap- t also **balances the misses**: 3.2% below and 2.1% above, against 4.7% and 1.3% for t .

