# Lecture 8: Robust Inference II

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

February 16, 2022

So far, we have relied on asymptotic limit quantities and limit distributions in our inferential procedures, relating test statistics to the normal or $t$ distribution:

- Number of sample units ($n$) goes to infinity.
- Number of clusters ($H$) goes to infinity.

So far, we have relied on asymptotic limit quantities and limit distributions in our inferential procedures, relating test statistics to the normal or $t$ distribution:

- Number of sample units ($n$) goes to infinity.
- Number of clusters ($H$) goes to infinity.

In some cases $n$ or $H$ may be too small for these asymptotic limits to characterize the distribution of our test statistics.

Particularly a problem if the data are "ill-behaved" (high maximal leverage – cf. Young, 2015a,b).

So far, we have relied on asymptotic limit quantities and limit distributions in our inferential procedures, relating test statistics to the normal or $t$ distribution:

- Number of sample units ($n$) goes to infinity.
- Number of clusters ($H$) goes to infinity.

In some cases $n$ or $H$ may be too small for these asymptotic limits to characterize the distribution of our test statistics.

Particularly a problem if the data are "ill-behaved" (high maximal leverage – cf. Young, 2015a,b).

Then, asymptotically valid standard errors may be too small and asymptotic test distributions may be highly inaccurate.

When this happens, we could be duped into thinking our estimates are more precise than they actually are given the size of our sample.

More formally,

- Our 95% confidence intervals would cover the true $\beta$ less than 95% of the time (error in coverage probability).
- Our null hypothesis tests would reject the null more often than $100\alpha\%$ of the time (error in rejection probability).

```
. reg income program presesirt, cluster(clu)

Linear regression                               Number of obs =       177
                                                F(  2,     3) =     12.83
                                                Prob > F      =    0.0339
                                                R-squared     =    0.0450
                                                Root MSE      =     38720

                                    (Std. Err. adjusted for 4 clusters in clus)
------------------------------------------------------------------------------
             |               Robust
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     program |   12462.21   2495.431     4.99   0.015     4520.631    20403.78
   presesirt |   1104.258   427.3683     2.58   0.082    -255.8186    2464.334
       _cons |   11279.24   3933.729     2.87   0.064    -1239.641    23798.12
------------------------------------------------------------------------------

. di tprob(3,2.58)
.08177981

. di tprob(174, 2.58)
.01070555

. di 2*(1-normprob(2.58))
.00988003
```

Analytical finite sample adjustments — e.g., Stata's cluster robust command uses (i) a degrees of freedom correction factor in computing the s.e. and (ii) tests against a $t$ distribution with $H-1$ rather than $n-k$ *df* — but these are based on approximations (for normal, iid data). See Young (2015b) for more on such corrections.

Estimators we consider are usually simple linear statistics (means, OLS coefficients, etc.) that have "asymptotically pivotal" test statistics with rather *simple* asymptotic distributions.[1]

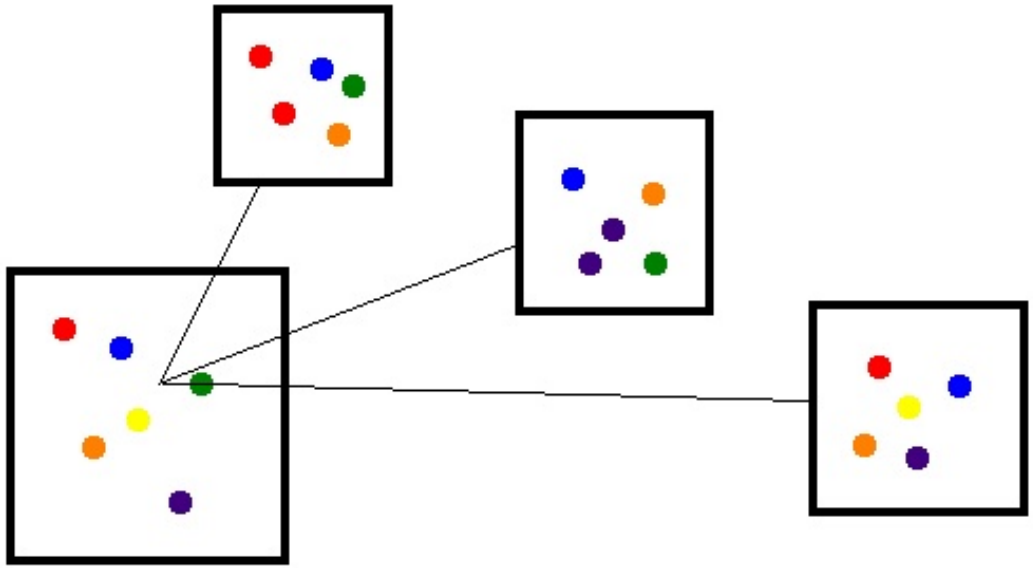
[1] An asymptotically pivotal statistic is one that has an asymptotic distribution that does not depend on unknown parameters governing the underlying data.

Estimators we consider are usually simple linear statistics (means, OLS coefficients, etc.) that have "asymptotically pivotal" test statistics with rather *simple* asymptotic distributions.[1]

Sometimes we might work with statistics with asymptotic distributions that are simply harder to characterize or with large first-order approximation error (e.g., ratio estimators, predicted values from non-linear models, complicated test statistics).

[1] An asymptotically pivotal statistic is one that has an asymptotic distribution that does not depend on unknown parameters governing the underlying data.

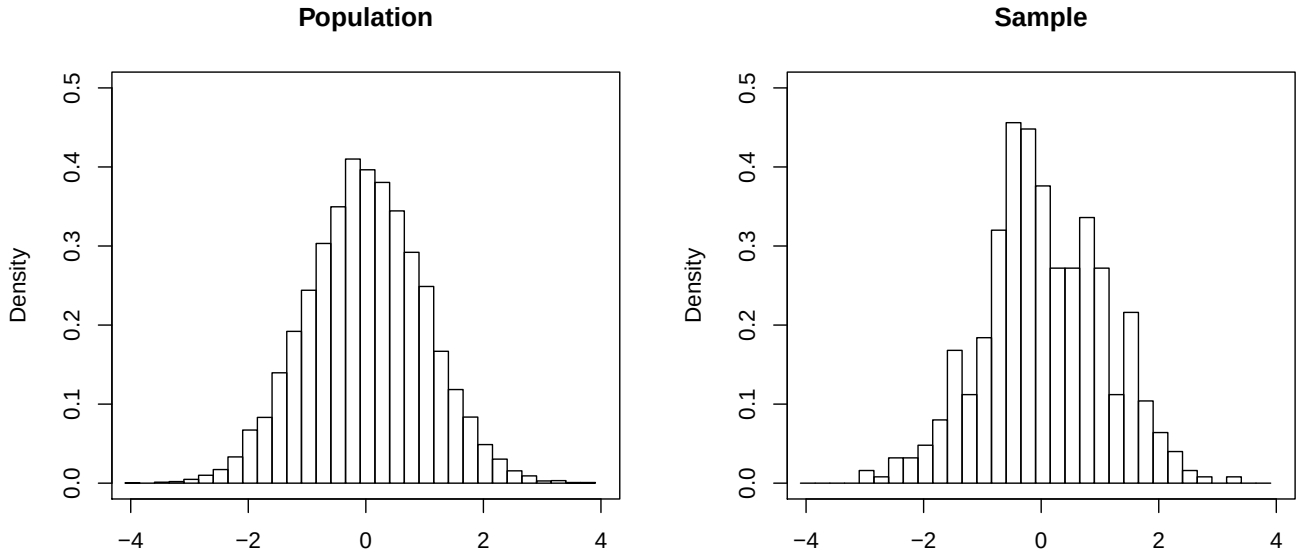
# Bootstrapping Overview



An approach to inference in such settings is bootstrapping:

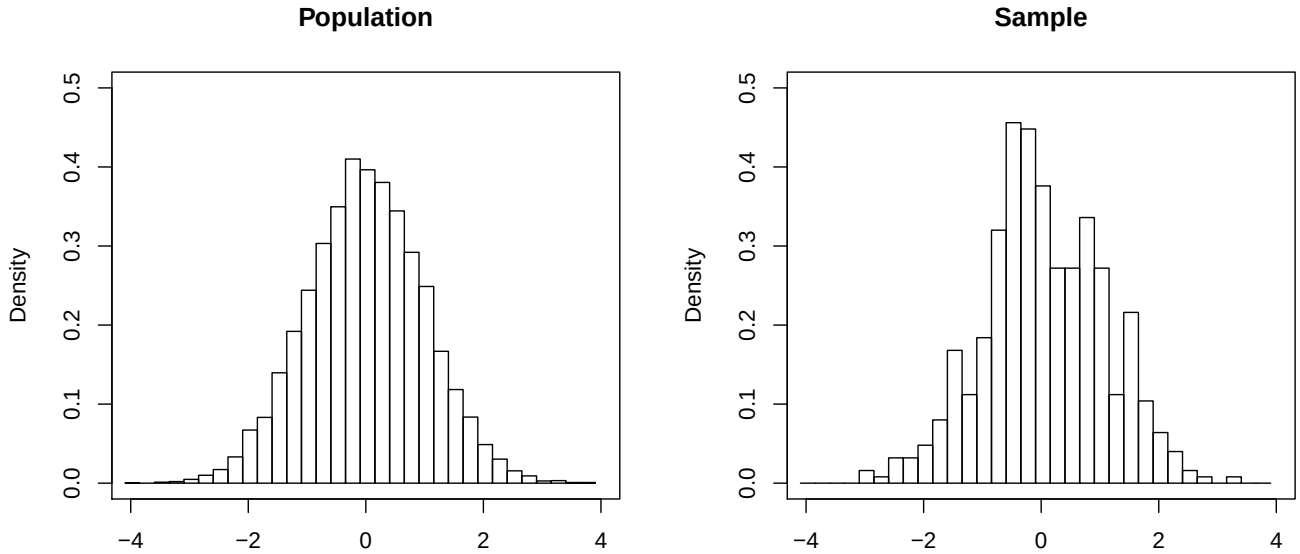
- Generate a large number of pseudo-samples (e.g., 1,000 of them) drawn from the original sample.
- Use these pseudo-samples to approximate the true sampling distribution.

# Bootstrapping Overview




Freedman (2009, Ch. 8) gives an intuitive introduction.

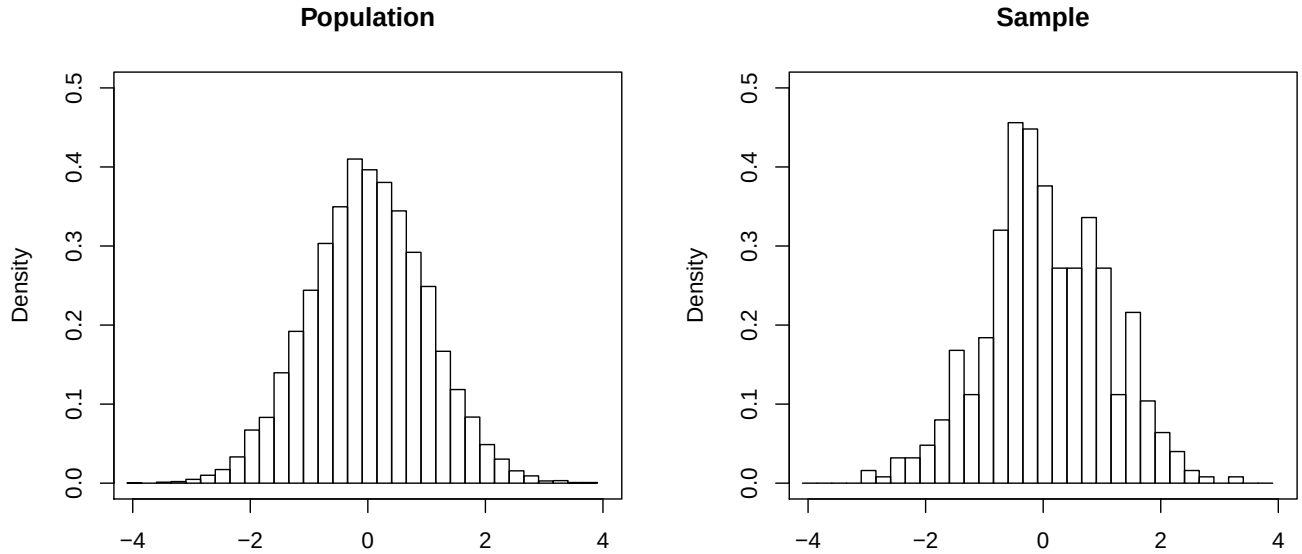
# Bootstrapping Overview




Freedman (2009, Ch. 8) gives an intuitive introduction.

- Random sample $S_n$ of size $n$ from large population, $P$.

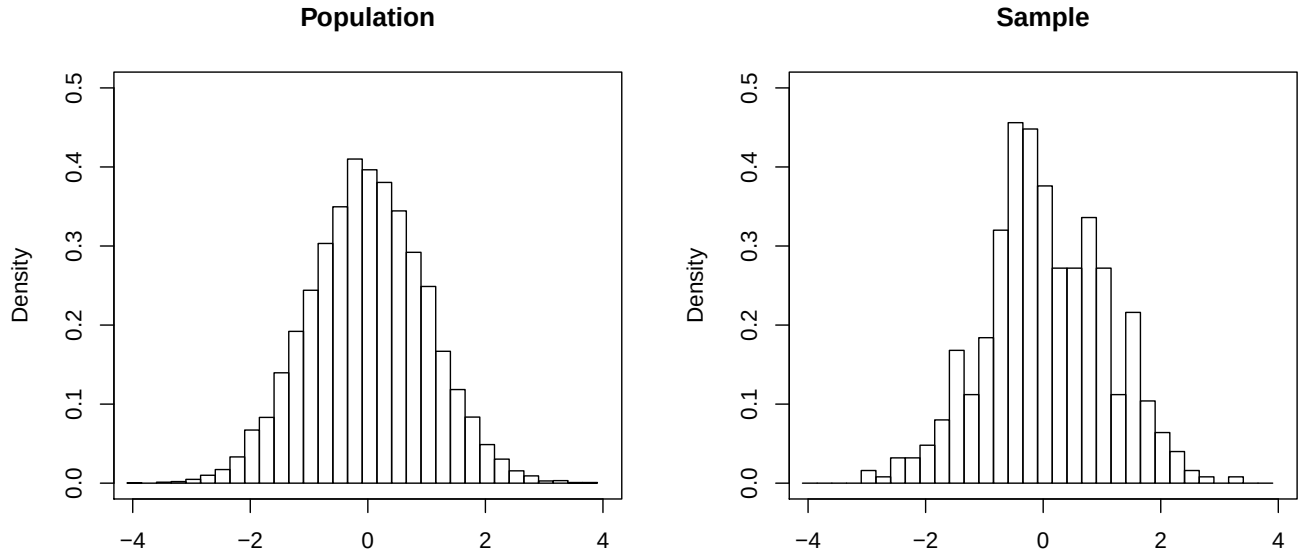
# Bootstrapping Overview




Freedman (2009, Ch. 8) gives an intuitive introduction.

- Random sample $S_n$ of size $n$ from large population, $P$.
- Distribution of variables in $S_n$ approximates distribution in $P$.

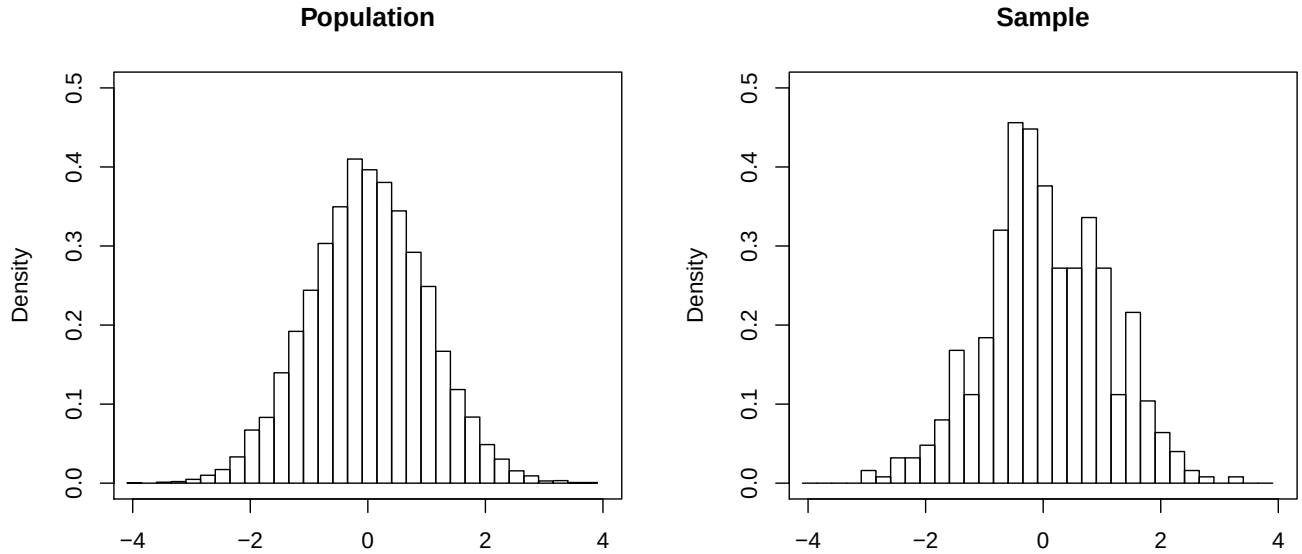
# Bootstrapping Overview




Freedman (2009, Ch. 8) gives an intuitive introduction.

- Random sample $S_n$ of size $n$ from large population, $P$.
- Distribution of variables in $S_n$ approximates distribution in $P$.
- This improves as $n \to \infty$.
- $\Rightarrow$ bootstrap validity is asymptotic.

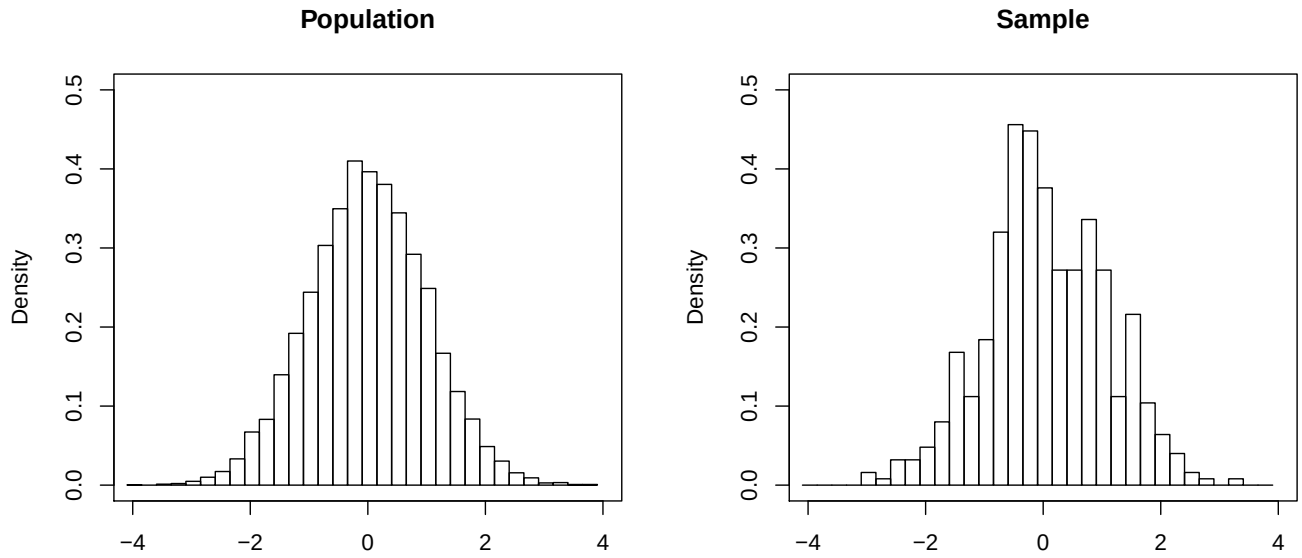
# Bootstrapping Overview




Freedman (2009, Ch. 8) gives an intuitive introduction.

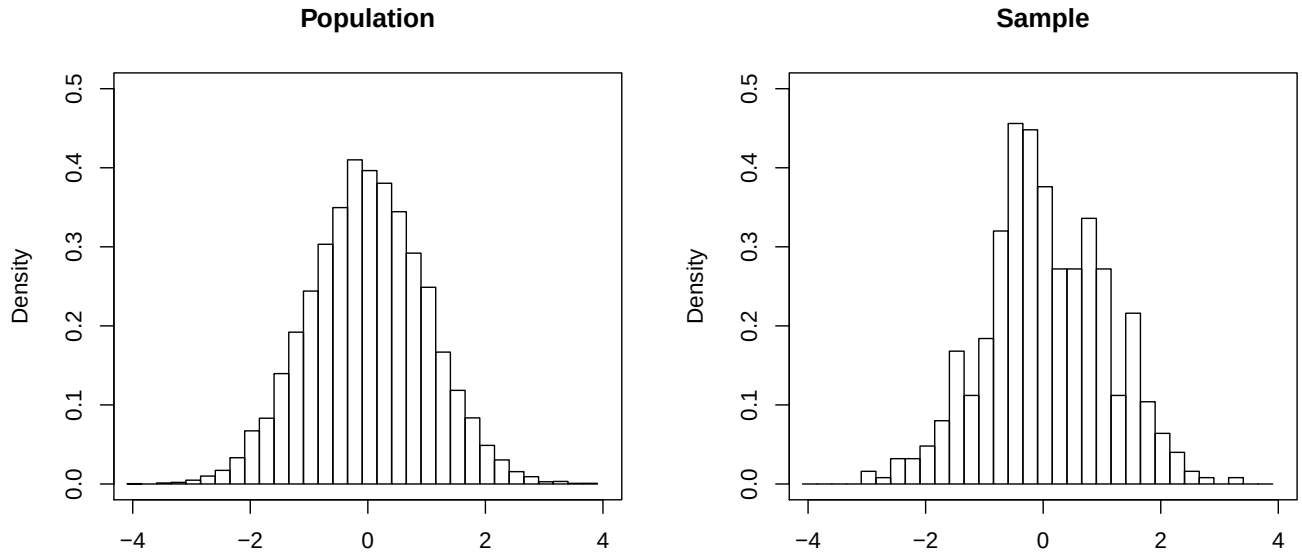
- Random sample $S_n$ of size $n$ from large population, $P$.
- Distribution of variables in $S_n$ approximates distribution in $P$.
- This improves as $n \to \infty$.
- $\Rightarrow$ bootstrap validity is asymptotic.
- Convergence of bootstrap variance estimators is often faster than those based on analytical (first order) approximations $\Rightarrow$ "asymptotic refinement".

# Bootstrapping Overview




- Sampling $n$ values *with replacement* from $S_n = \{1, \ldots, n\}$ approximates sampling $n$ values of $X$ from $P$.
- The distribution of statistics computed on the bootstrap samples drawn from $S_n$ approximates the distribution of statistics computed on samples from $P$.

# Bootstrapping Overview




In this case, $n = 500$ and $\sigma = 1$, and so the variance of the sample mean, $\bar{X}$ is $1/500 = 0.002$.

```
> ## Population ##
> n <- 10000
> set.seed(123)
> x <- rnorm(n)
> ## Sample ##
> n.s <- 500
> sampled.indices <- sample(1:n, n.s)
> x.s <- x[sample(sampled.indices, n.s)]
> ## X.bar for this sample ##
> mean(x.s)
[1] 0.04966862
> ## Bootstrap the mean ##
> n.boot <- 1000
> x.bar.b <- rep(NA,n.boot)
> for(i in 1:n.boot){
+  x.bar.b[i] <- mean(x[sample(sampled.indices,n.s, replace=T)])
+ }
> ## True mean and variance of X.bar ##
> 0;1/500
[1] 0
[1] 0.002
> ## Bootstrap mean and variance estimates ##
> mean(x.bar.b);var(x.bar.b)
[1] 0.04865665
[1] 0.00192289
```

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.
- Sample yields distribution $P_n$. Estimator $\hat{\theta}_n = f(P_n)$.

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.
- Sample yields distribution $P_n$. Estimator $\hat{\theta}_n = f(P_n)$.
- Define $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Suppose CDF of $R_n$ is $H_n$.

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.
- Sample yields distribution $P_n$. Estimator $\hat{\theta}_n = f(P_n)$.
- Define $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Suppose CDF of $R_n$ is $H_n$.
- If we knew $H_n$, could construct $100(1-\alpha)\%$ interval for $\hat{\theta}_n$.

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.
- Sample yields distribution $P_n$. Estimator $\hat{\theta}_n = f(P_n)$.
- Define $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Suppose CDF of $R_n$ is $H_n$.
- If we knew $H_n$, could construct $100(1-\alpha)\%$ interval for $\hat{\theta}_n$.
- We don't know $H_n$, so use bootstrap to approximate:

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.
- Sample yields distribution $P_n$. Estimator $\hat{\theta}_n = f(P_n)$.
- Define $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Suppose CDF of $R_n$ is $H_n$.
- If we knew $H_n$, could construct $100(1-\alpha)\%$ interval for $\hat{\theta}_n$.
- We don't know $H_n$, so use bootstrap to approximate:
  - Draw $B$ bootstrap samples. Estimate $\hat{\theta}_1^*, ..., \hat{\theta}_B^*$.
  - Then $\hat{H}_n(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(\sqrt{n}(\hat{\theta}_b^* - \hat{\theta}) \leq t\right)$.
  - Get $(\alpha/2, 1-\alpha/2)$ quantiles of $\hat{H}_n(t)$, call them $(\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2})$.
  - Bootstrap CI: $C_n = \left[\hat{\theta} - \frac{\hat{t}_{1-\alpha/2}}{\sqrt{n}}, \hat{\theta} - \frac{\hat{t}_{\alpha/2}}{\sqrt{n}}\right]$.
  - (Not a typo– $t_{\alpha/2}$ often negative. See Wasserman blog post.)

# Bootstrapping Overview

Bootstrap CI à la Wasserman (2006, *All of Nonparametric...* book):

- Population distribution $P$. Target functional $\theta = f(P)$.
- Sample yields distribution $P_n$. Estimator $\hat{\theta}_n = f(P_n)$.
- Define $R_n = \sqrt{n}(\hat{\theta}_n - \theta)$. Suppose CDF of $R_n$ is $H_n$.
- If we knew $H_n$, could construct $100(1-\alpha)\%$ interval for $\hat{\theta}_n$.
- We don't know $H_n$, so use bootstrap to approximate:
  - Draw $B$ bootstrap samples. Estimate $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
  - Then $\hat{H}_n(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}\left(\sqrt{n}(\hat{\theta}_b^* - \hat{\theta}) \leq t\right)$.
  - Get $(\alpha/2, 1-\alpha/2)$ quantiles of $\hat{H}_n(t)$, call them $(\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2})$.
  - Bootstrap CI: $C_n = \left[\hat{\theta} - \frac{\hat{t}_{1-\alpha/2}}{\sqrt{n}}, \hat{\theta} - \frac{\hat{t}_{\alpha/2}}{\sqrt{n}}\right]$.
  - (Not a typo– $t_{\alpha/2}$ often negative. See Wasserman blog post.)
- More refined version uses "studentized" quantities.
- Consistency of these bootstrap CIs depends on whether the limit for $H_n$ is continuous.

# Bootstrapping Overview

Other applications (continuing with notation from previous slide):

# Bootstrapping Overview

Other applications (continuing with notation from previous slide):

- Hypothesis tests:
  - Suppose we have $Y \sim P$, and want to test $H_0 : \mathrm{E}[Y] = 0$.
  - Test statistic: $\hat{t} = \frac{\bar{Y}}{\hat{\sigma}_{\tilde{Y}}/\sqrt{n}}$
  - Compute $\tilde{Y}_i = Y_i - \bar{Y}$ (this imposes $H_0$ on $P_n$).
  - Draw $B$ bootstrap and compute $\hat{t}_b^* = \frac{\bar{\tilde{Y}}_b^*}{\hat{\sigma}_{b,\tilde{Y}}^*/\sqrt{n}}$ for each.
  - Estimated $p$ value: $\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\hat{t}_b^*| \geq |\hat{t}|)$.

# Bootstrapping Overview

Other applications (continuing with notation from previous slide):

- Hypothesis tests:
  - Suppose we have $Y \sim P$, and want to test $H_0 : \mathrm{E}[Y] = 0$.
  - Test statistic: $\hat{t} = \frac{\bar{Y}}{\hat{\sigma}_{\tilde{Y}}/\sqrt{n}}$
  - Compute $\tilde{Y}_i = Y_i - \bar{Y}$ (this imposes $H_0$ on $P_n$).
  - Draw $B$ bootstrap and compute $\hat{t}_b^* = \frac{\bar{\tilde{Y}}_b^*}{\hat{\sigma}_{b,\tilde{Y}}^*/\sqrt{n}}$ for each.
  - Estimated $p$ value: $\frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\hat{t}_b^*| \geq |\hat{t}|)$.
- Bias correction:
  - If $P_n \approx P$, then $\mathrm{E}[\hat{\theta}^*] - \hat{\theta} \approx \mathrm{E}[\hat{\theta}_n] - \theta$.
  - Can use the LHS to correct bias in $\hat{\theta}$.

# Bootstrapping Overview

Bootstrapping has many flavors, based on:

- Observational units sampled–individual units, clusters.
- Objects sampled—$(Y, X)$ values, residuals.
- Statistic calculated with sample objects–parameter estimate, test statistic.
- Uses of the bootstrap distribution: variance estimation, estimate CI, hypothesis test, etc.

# Ratio estimators

- Last week we discussed inverse-propensity score weighting (IPSW) as an alternative to matching as a way to make use of the CIA assumption.

# Ratio estimators

- Last week we discussed inverse-propensity score weighting (IPSW) as an alternative to matching as a way to make use of the CIA assumption.
- The simple (unstabilized) IPSW estimator for the ATE takes the form,

$$\hat{\rho}_{IPSW} = \frac{1}{n}\sum_i \frac{D_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_i \frac{(1-D_i)Y_i}{1-\hat{e}(X_i)}.$$

  where $\hat{e}(X_i)$ is the estimated propensity score.
  (See Busso et al. for expressions for ATT, etc.)
- What is $\mathrm{Var}[\hat{\rho}_{IPSW}]$? What would be a good estimator for it?

# Ratio estimators

- When $\hat{e}(X_i)$ is estimated using a relatively simple model (viz., $M$-estimator with smooth estimating equations), first order asymptotic approximation of the variance is easy with Taylor expansion (aka "linearization" or "delta method").

# Ratio estimators

- When $\hat{e}(X_i)$ is estimated using a relatively simple model (viz., $M$-estimator with smooth estimating equations), first order asymptotic approximation of the variance is easy with Taylor expansion (aka "linearization" or "delta method").
- E.g., with logistic regression ($\hat{e}(X_i) = \Lambda(X_i\hat{\gamma})$):

$$\hat{\theta} = \begin{pmatrix} \hat{\gamma} \\ \hat{\rho}_{IPSW} \end{pmatrix} \text{ solves } \sum_i \psi(Y_i, X_i; \hat{\theta}) = \begin{pmatrix} \sum_i X_i \left[D_i - \Lambda(X_i\hat{\gamma})\right] \\ \sum_i \left[\frac{D_i Y_i}{\Lambda(X_i\hat{\gamma})} - \frac{(1-D_i)Y_i}{1-\Lambda(X_i\hat{\gamma})} - \hat{\rho}_{IPSW}\right] \end{pmatrix} = \mathbf{0}$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{MVN}\left(0, V_{\hat{\theta}}\right), \text{ with } V_{\hat{\theta}} = A^{-1}B\left(A^{-1}\right)'$$

and

$$A = \text{E}\left[-\frac{\partial}{\partial \theta'}\psi(Y_i, X_i, \theta_0)\right], B = \text{E}\left[\psi(Y_i, X_i, \theta_0)\psi(Y_i, X_i, \theta_0)'\right].$$

# Ratio estimators

- When $\hat{e}(X_i)$ is estimated using a relatively simple model (viz., $M$-estimator with smooth estimating equations), first order asymptotic approximation of the variance is easy with Taylor expansion (aka "linearization" or "delta method").
- E.g., with logistic regression ($\hat{e}(X_i) = \Lambda(X_i\hat{\gamma})$):

$$\hat{\theta} = \begin{pmatrix} \hat{\gamma} \\ \hat{\rho}_{IPSW} \end{pmatrix} \text{ solves } \sum_i \psi(Y_i, X_i; \hat{\theta}) = \begin{pmatrix} \sum_i X_i \left[ D_i - \Lambda(X_i\hat{\gamma}) \right] \\ \sum_i \left[ \frac{D_i Y_i}{\Lambda(X_i\hat{\gamma})} - \frac{(1-D_i)Y_i}{1-\Lambda(X_i\hat{\gamma})} - \hat{\rho}_{IPSW} \right] \end{pmatrix} = \mathbf{0}$$

and

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \text{MVN}\left(0, V_{\hat{\theta}}\right), \text{ with } V_{\hat{\theta}} = A^{-1} B \left(A^{-1}\right)'$$

and

$$A = \text{E}\left[ -\frac{\partial}{\partial \theta'} \psi(Y_i, X_i, \theta_0) \right], B = \text{E}\left[ \psi(Y_i, X_i, \theta_0) \psi(Y_i, X_i, \theta_0)' \right].$$

- A consistent sandwich estimator applies sample analogues (cf. Wooldridge $M$ estimation chapter).

# Ratio estimators

- But analytical solutions can sometimes be very difficult.
- Moreover, first-order approximation error may be large.

# Ratio estimators

- But analytical solutions can sometimes be very difficult.
- Moreover, first-order approximation error may be large.
- An approach is to use the bootstrap to estimate Var $[\hat{\rho}_{IPSW}]$: R example...

# Ratio estimators

- But analytical solutions can sometimes be very difficult.
- Moreover, first-order approximation error may be large.
- An approach is to use the bootstrap to estimate $\text{Var}\left[\hat{\rho}_{IPSW}\right]$:
  R example...
- Works when estimating equations (including for estimating the pscores) are asymptotically linear.
- Many possible refinements—e.g., working with bootstrap distributions of pivotal statistics (e.g., $t$ statistics), and then rescaling to obtain standard errors on the scale of the outcome variable.

  (cf. Efron & Tibshirani, 1993; Horowitz, 2001, 2003)

# Bootstrap consistency

- Horowitz (2001) reviews formal conditions for the bootstrap to be consistent in characterizing the sampling distribution.

# Bootstrap consistency

- Horowitz (2001) reviews formal conditions for the bootstrap to be consistent in characterizing the sampling distribution.
- Essentially *smoothness* conditions (Beran & Ducharme 1991):
  - Estimator of interest must be smooth in perturbations of the sample values
  - Limiting distribution of the estimator must be smooth in perturbations of the estimator value.
- With linear-functional estimators (i.e., estimators that we can write as scaled sums of transformed variables, like means, regression coefficients, etc.), asymptotic normality necessary and sufficient for bootstrap consistency (Mammen 1992).

# Application: Bootstrap Inference with Clustered Data

# Application: Inference with Clustered Data

Another application relevant to our discussion last week is presented in Cameron et al. (2008), who review bootstrap methods for clustered data when number of clusters ($H$) is small.

They compare asymptotically valid cluster robust variance estimators to a variety of bootstrap estimators for clustered data.

# Application: Inference with Clustered Data

Pairs cluster bootstrap-s.e. (a.k.a. "block bootstrap"): sample $H$ clusters of $(Y,X)$ values with replacement, compute $\hat{\beta}_b^*$. Use $s.d.(\hat{\beta}_b^*)$ as s.e. estimate. Use usual tests, although perhaps against a modified distribution (e.g., use $t$ with $H-1$ $df$, like Stata with cluster-robust).

# Application: Inference with Clustered Data

Pairs cluster bootstrap-$t$: sample $H$ clusters of $(Y,X)$ values with replacement, compute $w_b^*(\theta) = (\hat{\beta}_b^* - \theta)/\widehat{s.e.}_{CR}(\hat{\beta}_b^*)$. For confidence intervals, use,

$$\hat{\beta} \pm t^*_{1-\alpha/2}\widehat{s.e.}_{CR}(\hat{\beta}),$$

where $t^*_{1-\alpha/2}$ is the appropriate quantile of the distribution of $w_b^*(\hat{\beta})$. For a two-sided $p$-value, use proportion of $w_b^*(0)$ values that are at least $|\hat{\beta}/\widehat{s.e.}_{CR}(\hat{\beta})|$ from zero.

# Application: Inference with Clustered Data

Some issues arise for these pair clusters bootstrap methods:

- For pair clusters bootstrap-s.e., using "usual tests" means no asymptotic refinement—does not account for irregularity of small sample distributions.
- Pair clusters bootstrap-$t$ provides asymptotic refinement to account for this.
- When treatment is binary and $H$ is small, bootstrap samples may yield all 0's or all 1's, making $\hat{\beta}_b^*$ inestimable. Would be better if we could avoid this.

# Application: Inference with Clustered Data

Residual cluster bootstrap (s.e. or $t$): sample $H$ clusters of residuals with replacement to form $\{\hat{e}_1^*, ..., \hat{e}_H^*\}$, then apply these to the $H$ clusters of $X\hat{\beta}$ values to construct $\{(\hat{y}_1^*, X_1)..., (\hat{y}_H^*, X_H)\}$. Compute $\hat{\beta}_b^*$, and perform either bootstrap or bootstrap-$t$ inference with each bootstrap sample.

For testing the null hypothesis, use residuals from the regression that imposes the null, and construct the $y_h^*$ values with the null imposed. Then proceed as above.

Problem: based on assumption that $e_h$ vectors are iid from cluster to cluster, and presumes that clusters are all equal size, which are not things that cluster-robust standard errors assume.

# Application: Inference with Clustered Data

Wild cluster bootstrap (s.e. or $t$): form $\{\hat{e}_1^*, ..., \hat{e}_H^*\}$ by letting $\hat{e}_h^* = \hat{e}_h$ with probability 0.5 and $\hat{e}_h^* = -\hat{e}_h$ with probability 0.5 (justified for symmetric error distributions). Then apply these to the $H$ clusters of $X\hat{\beta}$ values to construct $\{(\hat{y}_1^*, X_1)..., (\hat{y}_H^*, X_H)\}$. Compute $\hat{\beta}_b^*$, and perform either bootstrap or bootstrap-$t$ inference with each bootstrap sample.

Again, for testing the null hypothesis, use residuals from the regression that imposes the null, and construct the $y_h^*$ values with the null imposed. Then proceed as above.

(See MacKinnon & Webb 2014 and MacKinnon 2015 for a current discussion.)

# Application: Inference with Clustered Data

TABLE 4.—1,000 SIMULATIONS FROM DIFFERENT DGPS (SEE TEXT) AND $G = 10$ Groups
(Rejection rates for tests of nominal size 0.05 with simulation standard errors in parentheses)

| Estimator # | Method | Column Number | Main—from Table 2 | Reject based on T (8 dof) | Cluster Size = 2 | Cluster Size = 10 | Cluster Size = 100 | 4 RHS Variables | Xs are Constant Within Group | Xs Are i.i.d. | Unbalanced Group Sizes (10, 50) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 1 | Assume i.i.d. | | 0.491 (0.016) | | 0.106 (0.010) | 0.268 (0.014) | 0.679 (0.015) | 0.687 (0.015) | 0.770 (0.013) | 0.054 (0.007) | 0.524 (0.016) |
| 2 | Moulton-type estimator | | 0.092 (0.009) | 0.044 (0.006) | 0.095 (0.009) | 0.098 (0.009) | 0.088 (0.009) | 0.089 (0.009) | 0.125 (0.010) | 0.061 (0.008) | 0.129 (0.011) |
| 3 | Cluster-robust | | 0.129 (0.010) | 0.082 (0.009) | 0.137 (0.010) | 0.126 (0.010) | 0.115 (0.010) | 0.129 (0.010) | 0.183 (0.013) | 0.103 (0.010) | 0.183 (0.012) |
| 4 | CR3 residual correction | | 0.090 (0.009) | 0.054 (0.007) | 0.094 (0.009) | 0.086 (0.009) | 0.077 (0.008) | 0.080 (0.009) | 0.090 (0.009) | 0.086 (0.009) | 0.091 (0.009) |
| 5 | Pairs cluster bootstrap-se | | 0.120 (0.010) | 0.071 (0.008) | 0.100 (0.009) | 0.114 (0.010) | 0.120 (0.010) | 0.128 (0.010) | 0.063 (0.008) | 0.122 (0.010) | 0.138 (0.011) |
| 6 | Residual cluster bootstrap-se | | 0.058 (0.007) | 0.013 (0.004) | 0.069 (0.008) | 0.068 (0.008) | 0.060 (0.008) | 0.057 (0.007) | 0.054 (0.007) | 0.080 (0.009) | |
| 7 | Wild cluster bootstrap-se | | 0.028 (0.005) | 0.006 (0.002) | 0.048 (0.007) | 0.044 (0.006) | 0.032 (0.006) | 0.030 (0.005) | 0.036 (0.006) | 0.053 (0.007) | 0.019 (0.004) |
| 8 | Pairs cluster bootstrap-BCA | | 0.111 (0.010) | | 0.125 (0.010) | 0.112 (0.010) | 0.109 (0.010) | 0.112 (0.010) | 0.100 (0.009) | 0.134 (0.011) | 0.140 (0.011) |
| 9 | BDM bootstrap-t | | 0.119 (0.010) | | 0.086 (0.009) | 0.115 (0.010) | 0.112 (0.010) | 0.119 (0.010) | 0.121 (0.010) | 0.097 (0.009) | 0.128 (0.011) |
| 10 | Pairs cluster bootstrap-t | | 0.096 (0.009) | | 0.085 (0.009) | 0.083 (0.009) | 0.086 (0.009) | 0.090 (0.009) | 0.066 (0.008) | 0.079 (0.009) | 0.120 (0.010) |
| 11 | Pairs CR3 bootstrap-t | | 0.090 (0.009) | | 0.075 (0.008) | 0.077 (0.008) | 0.081 (0.009) | 0.084 (0.009) | 0.050 (0.007) | 0.082 (0.009) | 0.110 (0.010) |
| 12 | Residual cluster bootstrap-t | | 0.055 (0.007) | | 0.052 (0.007) | 0.056 (0.007) | 0.050 (0.007) | 0.043 (0.006) | 0.043 (0.006) | 0.065 (0.008) | |
| 13 | Wild cluster bootstrap-t | | 0.055 (0.007) | | 0.064 (0.008) | 0.056 (0.007) | 0.048 (0.007) | 0.052 (0.007) | 0.045 (0.007) | 0.064 (0.008) | 0.061 (0.008) |
| | T_distribution(8) | | 0.086 | | | | | | | | |

# Remarks: Bootstrap failure

- The bootstrap depends on large sample size or low maximal leverage, although in a manner that is less sensitive than conventional "robust" analytical approximations (Young 2015a).

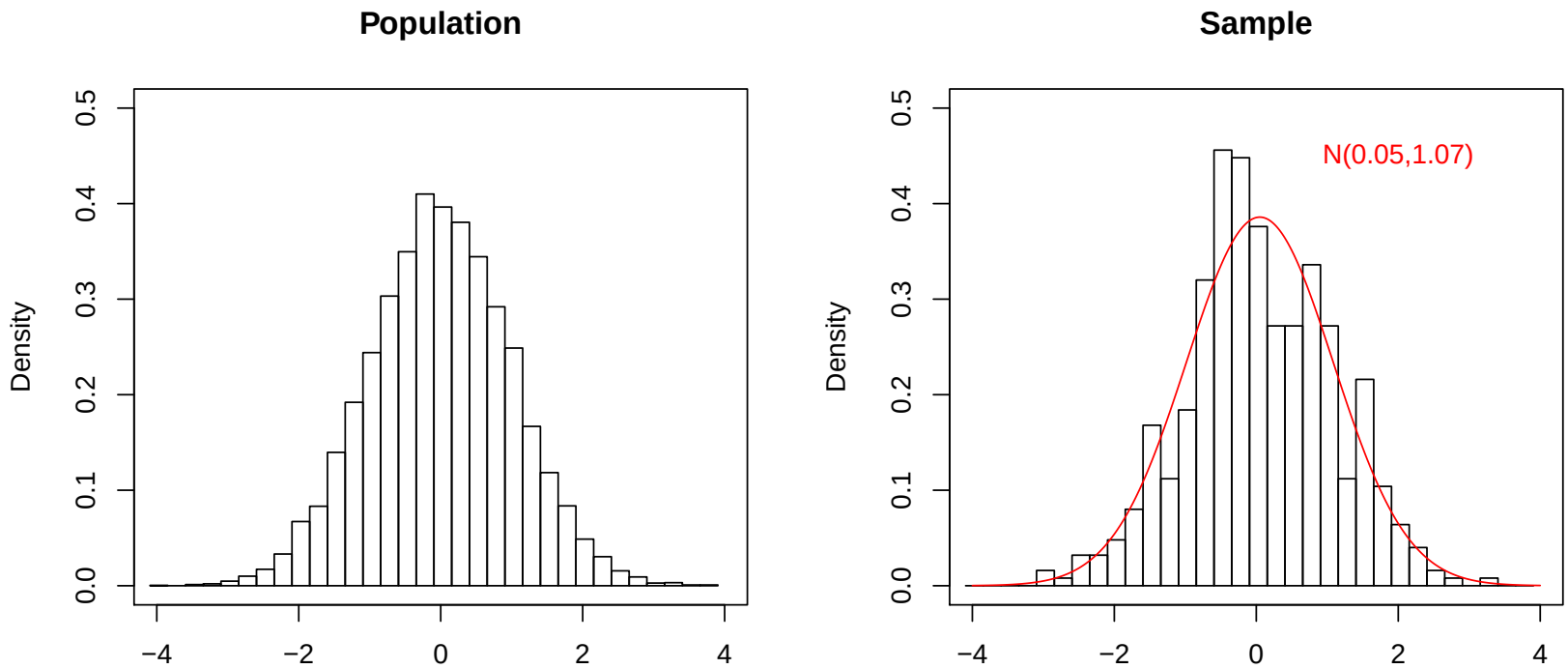
# Remarks: Bootstrap failure

- The bootstrap depends on large sample size or low maximal leverage, although in a manner that is less sensitive than conventional "robust" analytical approximations (Young 2015a).
- The bootstrap fails when sampling distribution of estimator is not smooth.
  - Example: estimates of extrema, e.g., maximum: R example

# Remarks: Bootstrap failure

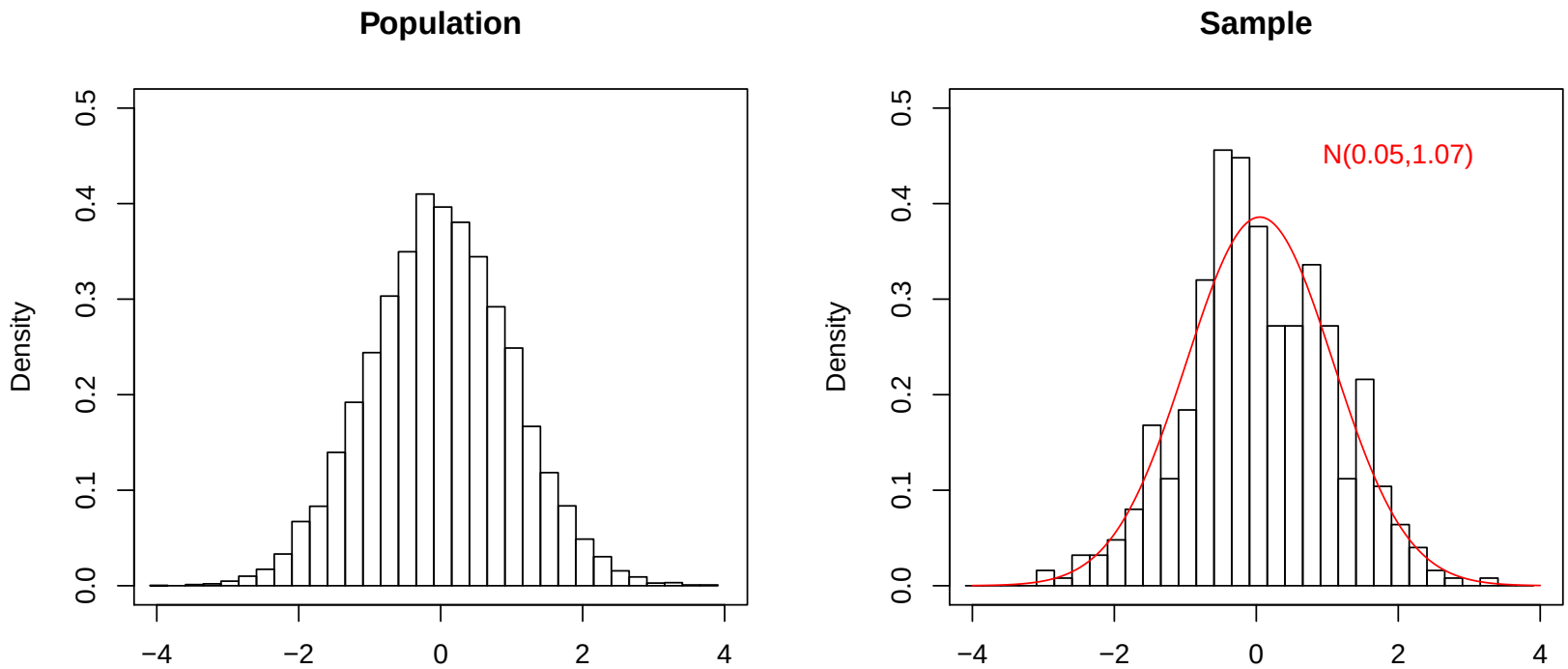
- The bootstrap depends on large sample size or low maximal leverage, although in a manner that is less sensitive than conventional "robust" analytical approximations (Young 2015a).
- The bootstrap fails when sampling distribution of estimator is not smooth.
  - Example: estimates of extrema, e.g., maximum: R example
  - Abadie and Imbens (2008) review bootstrapping for unconditional variance estimation for matching estimators:
    - Ordinary bootstrapping fails for nearest-neighbor matching due to such non-smoothness problems. They derive a consistent analytic variance estimator instead.
    - They conjecture that other types of bootstrapping may be okay (e.g., subsampling from exact-match stratification cells or bootstrapped kernel-weighting matching).
    - See Bodory et al. (2016) for an update.

# Remarks: Parametric bootstrap




Examples we have considered here are all non-parametric, which involves approximating the population distribution with the *empirical* distribution of the sample.

# Remarks: Parametric bootstrap




Examples we have considered here are all non-parametric, which involves approximating the population distribution with the *empirical* distribution of the sample. Parametric bootstrap draws from a parametric distribution fit to the observed data (e.g., `Clarify` in Stata or `zelig` in R).

# Remarks: Bootstrap vs. permutation tests

Bootstrapping under the null hypothesis resembles the exact permutation (or randomization) tests of the "sharp null" that we briefly discussed in lecture 2.

# Remarks: Bootstrap vs. permutation tests

Bootstrapping under the null hypothesis resembles the exact permutation (or randomization) tests of the "sharp null" that we briefly discussed in lecture 2.

- There, we (i) filled in potential outcomes under the sharp null, (ii) permuted or resampled treatment assignments, (iii) computed a test statistic, $t_b$, and then (iv) computed a $p$-value for the $t$ from the sample using the distribution of $t_b$.

# Remarks: Bootstrap vs. permutation tests

Bootstrapping under the null hypothesis resembles the exact permutation (or randomization) tests of the "sharp null" that we briefly discussed in lecture 2.

- There, we (i) filled in potential outcomes under the sharp null, (ii) permuted or resampled treatment assignments, (iii) computed a test statistic, $t_b$, and then (iv) computed a $p$-value for the $t$ from the sample using the distribution of $t_b$.
- This *isolates* the variability due to randomization.

# Remarks: Bootstrap vs. permutation tests

Bootstrapping under the null hypothesis resembles the exact permutation (or randomization) tests of the "sharp null" that we briefly discussed in lecture 2.

- There, we (i) filled in potential outcomes under the sharp null, (ii) permuted or resampled treatment assignments, (iii) computed a test statistic, $t_b$, and then (iv) computed a $p$-value for the $t$ from the sample using the distribution of $t_b$.
- This *isolates* the variability due to randomization.
- In our bootstrap examples we either resample whole units or hold the treatment and covariate distribution in the sample fixed and resample outcomes. In either case, bootstrapping approximates variability due to the *combination* of sampling and randomization.

# Remarks: Bootstrap vs. permutation tests

Bootstrapping under the null hypothesis resembles the exact permutation (or randomization) tests of the "sharp null" that we briefly discussed in lecture 2.

- There, we (i) filled in potential outcomes under the sharp null, (ii) permuted or resampled treatment assignments, (iii) computed a test statistic, $t_b$, and then (iv) computed a $p$-value for the $t$ from the sample using the distribution of $t_b$.
- This *isolates* the variability due to randomization.
- In our bootstrap examples we either resample whole units or hold the treatment and covariate distribution in the sample fixed and resample outcomes. In either case, bootstrapping approximates variability due to the *combination* of sampling and randomization.
- See Morgan (2017) for a discussion of how bootstrap and permutation tests compare for inference with experiments.

# Remarks: Bootstrap vs. permutation tests

- For randomized experiments, exact tests are robust (under the null, they are *exact* even in finite samples) and often more powerful for testing the sharp null relative to bootstrap or analytical approximations (Imbens & Rubin, 2015, Ch. 5; Young, 2015a).

# Remarks: Bootstrap vs. permutation tests

- For randomized experiments, exact tests are robust (under the null, they are *exact* even in finite samples) and often more powerful for testing the sharp null relative to bootstrap or analytical approximations (Imbens & Rubin, 2015, Ch. 5; Young, 2015a).
- Power of exact tests can be affected by choice of test statistics—$t$ stats, ranks, etc. Most available results assume the null.

# Remarks: Bootstrap vs. permutation tests

- For randomized experiments, exact tests are robust (under the null, they are *exact* even in finite samples) and often more powerful for testing the sharp null relative to bootstrap or analytical approximations (Imbens & Rubin, 2015, Ch. 5; Young, 2015a).
- Power of exact tests can be affected by choice of test statistics—$t$ stats, ranks, etc. Most available results assume the null.
- Chung and Romano (2013) discuss robustness of studentized tests (e.g., using $t$-stats and other pivotal statistics), which is similar to recommended practice for bootstrap.

# Remarks: Bootstrap vs. permutation tests

- For randomized experiments, exact tests are robust (under the null, they are *exact* even in finite samples) and often more powerful for testing the sharp null relative to bootstrap or analytical approximations (Imbens & Rubin, 2015, Ch. 5; Young, 2015a).
- Power of exact tests can be affected by choice of test statistics—$t$ stats, ranks, etc. Most available results assume the null.
- Chung and Romano (2013) discuss robustness of studentized tests (e.g., using $t$-stats and other pivotal statistics), which is similar to recommended practice for bootstrap.
- For experimental studies, the recommendation is to always report permutation test $p$-values (based on either $t$ or Wald statistics) as a basic check on whether there are any effects.

# Remarks: Bootstrap vs. permutation tests

- For randomized experiments, exact tests are robust (under the null, they are *exact* even in finite samples) and often more powerful for testing the sharp null relative to bootstrap or analytical approximations (Imbens & Rubin, 2015, Ch. 5; Young, 2015a).
- Power of exact tests can be affected by choice of test statistics—$t$ stats, ranks, etc. Most available results assume the null.
- Chung and Romano (2013) discuss robustness of studentized tests (e.g., using $t$-stats and other pivotal statistics), which is similar to recommended practice for bootstrap.
- For experimental studies, the recommendation is to always report permutation test $p$-values (based on either $t$ or Wald statistics) as a basic check on whether there are any effects.
- Such procedures can be applied in observational studies under an assumption of conditional random assignment (Rosenbaum, 2002).