Chapter 4
Replication Variance Estimation
Create a new sample by deleting one observation

\[
\bar{x}^{(k)} = \frac{n\bar{x} - x_k}{n-1} \quad \text{and} \quad \bar{x}^{(k)} - \bar{x} = -\frac{x_k - \bar{x}}{n-1}
\]

\[
\therefore \frac{n-1}{n} \sum_{k=1}^{n} (\bar{x}^{(k)} - \bar{x})^2 = \frac{1}{n(n-1)} \sum_{k=1}^{n} (x_k - \bar{x})^2 = n^{-1}s_x^2
\]
Alternative Jackknife Weights

\[
\tilde{x}^{(k)}_\psi = \psi x_k + (1 - \psi)\bar{x}^{(k)}
\]

\[
\tilde{x}^{(k)}_\psi - \bar{x} = \psi (x_k - \bar{x}) + (1 - \psi) (\bar{x}^{(k)} - \bar{x})
\]

\[
\tilde{x}^{(k)}_\psi - \bar{x} = (\psi - \frac{1 - \psi}{n - 1}) (x_k - \bar{x}) = (\frac{n\psi - 1}{n - 1}) (x_k - \bar{x})
\]

\[
\sum_{k=1}^{n} (\tilde{x}^{(k)}_\psi - \bar{x})^2 = \frac{(n\psi - 1)^2}{(n - 1)^2} \sum_{k=1}^{n} (x_k - \bar{x})^2
\]

If \((n\psi - 1)^2 = \frac{n - 1}{n} (1 - f)\), then \(\sum_{k=1}^{n} (\tilde{x}^{(k)}_\psi - \bar{x})^2 = \frac{1}{n} (1 - f) s_x^2\).
Random Group Jackknife

\( n = mb \) : m groups of size b, \( \bar{x}_1, \ldots, \bar{x}_m \)

\[
\bar{x} = \frac{1}{m} \sum_{i=1}^{m} \bar{x}_i
\]

\[
\hat{V}(\bar{x}) = \frac{1}{m} \frac{1}{m-1} \sum_{i=1}^{m} (\bar{x}_i - \bar{x})^2
\]

\[
\bar{x}_n^{(k)} = \frac{n\bar{x} - b\bar{x}_k}{n-b} = \frac{m\bar{x} - \bar{x}_k}{m-1}
\]

\[
\bar{x}_b^{(k)} - \bar{x} = \frac{n\bar{x} - b\bar{x}_k}{n-b} - \bar{x} = -\frac{1}{m-1}(\bar{x}_k - \bar{x})
\]

\[
\hat{V}_{RGJK}(\bar{x}) \equiv \frac{m-1}{m} \sum_{k=1}^{m} (\bar{x}_b^{(k)} - \bar{x})^2 = \frac{1}{m(m-1)} \sum_{k=1}^{m} (\bar{x}_k - \bar{x})^2
\]

Unbiased but d.f. = \( m - 1 \).
Theorem 4.1.1

\[ \mathcal{F}_N = \{y_1, \ldots, y_N\} : \text{sequence of finite population} \]
\[ y_i \sim iid(\mu_y, \sigma_y^2) \text{ with finite } 4 + \delta \text{ moments} \]
\[ g(\cdot) : \text{continuous function with continuous first derivative at } \mu_y \]

\[ \Rightarrow \frac{n - 1}{n} \sum_{k=1}^{n} \{g(\bar{y}^{(k)}) - g(\bar{y})\}^2 = [g'(\bar{y})]^2 \hat{V}(\bar{y}) + o_p(n^{-1}) \]

where \( \hat{V}(\bar{y}) = n^{-1}s^2 \) and \( g'(\bar{y}) = \frac{\partial g(\bar{y})}{\partial \bar{y}} \).
Proof of Theorem 4.1.1

By a Taylor linearization,

\[ g(\bar{y}^{(k)}) = g(\bar{y}) + g'(\bar{y}_k^*)(\bar{y}^{(k)} - \bar{y}) = g(\bar{y}) + g'(\bar{y})(\bar{y}^{(k)} - \bar{y}) + R_{nk}(\bar{y}^{(k)} - \bar{y}) \]

for some \( \bar{y}_k^* \in B_{\delta_k}(\bar{y}) \), \( \delta_k = \| \bar{y}^{(k)} - \bar{y} \| \) where \( R_{nk} = g'(\bar{y}_k^*) - g'(\bar{y}) \). Thus,

\[
\sum_{k=1}^{n} \left\{ g(\bar{y}^{(k)}) - g(\bar{y}) \right\}^2 = \sum_{k=1}^{n} \left\{ g'(\bar{y}_k^*) \right\}^2 (\bar{y}^{(k)} - \bar{y})^2
\]

\[
= \sum_{k=1}^{n} \left\{ g'(\bar{y}) + R_{nk} \right\}^2 (\bar{y}^{(k)} - \bar{y})^2
\]

(i) \( \max_{1 \leq k \leq n} |\bar{y}_k^* - \bar{y}| \to 0 \) in probability.

(ii) \( \max_{1 \leq k \leq n} |g'(\bar{y}_k^*) - g'(\bar{y})| \to 0 \) in probability.

\[ \therefore \quad \frac{n-1}{n} \sum_{k=1}^{n} \left\{ g(\bar{y}^{(k)}) - g(\bar{y}) \right\}^2 = [g'(\bar{y})]^2 \hat{V}(\bar{y}) + o_p(n^{-1}) \]
Proof of (i)

\[
\Pr[\max_k |\bar{y}^{(k)} - \bar{y}| > \epsilon] \leq \sum_{k=1}^{n} \Pr[|\bar{y}^{(k)} - \bar{y}| > \epsilon] \\
\leq \epsilon^{-2} \sum_{k=1}^{n} E[|\bar{y}^{(k)} - \bar{y}|^2] \\
\leq cV(\bar{y}) \rightarrow 0
\]

for some \( c \).

\[\therefore \max_k |\bar{y}^{(k)} - \bar{y}| \rightarrow 0 \text{ in probability.}\]

\[\therefore \max_k |\bar{y}^*_k - \bar{y}| \leq \max_k |\bar{y}^{(k)} - \bar{y}| \rightarrow 0\]
Proof of (ii)

Since $g'(\bar{y}) \to g'(\mu_y)$ in probability, it suffices to show that

$$\max_k |g'(\bar{y}_k^*) - g'(\mu_y)| = o_p(1).$$

Now, note that

$$Pr \left\{ \max_k |g'(\bar{y}_k^*) - g'(\mu_y)| > \epsilon \right\} \leq Pr \{ \max_k |\bar{y}_k^* - \mu_y| > \delta \}$$

$$+ Pr \left\{ \max_k |\bar{y}_k^* - \mu_y| < \delta, \max_k |g'(\bar{y}_k^*) - g'(\mu_y)| > \epsilon \right\}$$

holds for any $\delta > 0$. Since $g'(y)$ is continuous at $y = \mu_y$, we can choose $\delta$ such that the second term equals to zero. The first term converges to zero by (i).
If \( g(\cdot) \) has continuous second derivatives at \( \mu_y \), then

\[
\frac{n^{-1}(n-1)}{\sum_{k=1}^{n}[g(\bar{y}^{(k)}) - g(\bar{y})]^2 = [g'(\bar{y})]^2 \hat{V}(\bar{y}) + O_p(n^{-2})}.
\]

**Proof:**

\[
[g(\bar{y}^{(k)}) - g(\bar{y})]^2 = [g'(\bar{y}_k^*)]^2(\bar{y}^{(k)} - \bar{y})^2
\]

\[
g'(\bar{y}_k^*)^2 = [g'(\bar{y})]^2 + 2[g'(\bar{y}^{**})]g''(\bar{y}^{**})(\bar{y}_k^* - \bar{y})
\]

\[
\Rightarrow [g'(\bar{y}_k^*)]^2 = [g'(\bar{y})]^2 + K_1|\bar{y}_k^* - \bar{y}|
\]

for some \( K_1 \). Thus, since \( \bar{y}_k^* - \bar{y} = O_p(n^{-1}) \), we have the result.
Jackknife Often Larger than Taylor

\[ \hat{R} = \bar{x}^{-1}\bar{y} \]

\[ \hat{R}^{(k)} - \hat{R} = [\bar{x}^{(k)}]^{-1}\bar{y}^{(k)} - \bar{x}^{-1}\bar{y} \]

\[ = -[\bar{x}^{(k)}]^{-1}(y_k - \hat{R}x_k)(n - 1)^{-1} \]

\[ \hat{V}_{JK}(\hat{R}) = \frac{n-1}{n} \sum_{k=1}^{n} (\hat{R}^{(k)} - \hat{R})^2 \]

\[ = \frac{1}{n(n-1)} \sum_{k=1}^{n} [\bar{x}^{(k)}]^{-2}(y_k - \hat{R}x_k)^2 \]

vs. \[ \hat{V}_L(\hat{R}) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (\bar{x})^{-2}(y_k - \hat{R}x_k)^2 \]

\[ E[(\bar{x}^{(k)})^{-2}] \geq E[(\bar{x})^{-2}] \]
Quantiles

$$\xi_p = Q(p) = F^{-1}(p), \ p \in (0, 1) \text{ where } F(y) \text{ is cdf }$$

$$\hat{\xi}_p = \hat{Q}(p) = \inf \{ \hat{F}(y) \geq p \}, \ p = 0.5 \text{ for median }$$

$$\hat{F}(y) = \left( \sum_{i \in A} w_i \right)^{-1} \sum_{i \in A} w_i I(y_i \leq y)$$

To reduce the bias, use interpolation:

$$\hat{\xi}_p = \hat{\xi}_{p0} + \frac{\hat{\xi}_{p1} - \hat{\xi}_{p0}}{\hat{F}(\hat{\xi}_{p1}) - \hat{F}(\hat{\xi}_{p0})} \{ p - \hat{F}(\hat{\xi}_{p0}) \}$$

where $$\hat{\xi}_{p1} = \inf_{x_1, \ldots, x_n} \{ x; \hat{F}(x) \geq p \}$$ & $$\hat{\xi}_{p0} = \sup_{x_1, \ldots, x_n} \{ x; \hat{F}(x) \leq p \}$$
Test Inversion for Quantile C.I.

Construct acceptance region for $H_0 : p = p_0$

$$\{ p_0 - 2[\hat{V}(\hat{p}_0)]^{1/2}, p_0 + 2[\hat{V}(\hat{p}_0)]^{1/2} \}$$

Invert $p$-interval to give C.I. for $\xi_{p_0}$

$$\{ \hat{Q}(p_0 - 2[\hat{V}(\hat{p}_0)]^{1/2}), \hat{Q}(p_0 + 2[\hat{V}(\hat{p}_0)]^{1/2}) \}$$
Figure: Plot of CDF
Figure: Inverse CDF
Bahadur representation

- Let \( \hat{F}(x) \) be an unbiased estimator of \( F(x) \), the population CDF of \( X \).
- For given \( p \in (0, 1) \), we can define \( \zeta_p = F^{-1}(p) \) to be the \( p \)-th population quantile of \( X \).
- Let \( \hat{\zeta}_p = \hat{F}^{-1}(p) \) be the sample estimator of \( \zeta_p \) using \( \hat{F} \). Also, define \( \hat{p} = \hat{F}(\zeta_p) \). We are interested in applying a Taylor linearization of \( \hat{\zeta}_p \) around \( p = \hat{p} \).
Note that, by Taylor linearization,

\[ \zeta_p = \hat{F}^{-1}(\hat{p}) \]

\[ \approx \hat{F}^{-1}(p) + \frac{d\hat{F}^{-1}(p)}{dp}(\hat{p} - p) \]

\[ \approx \hat{F}^{-1}(p) + \frac{dF^{-1}(p)}{dp}(\hat{p} - p) \]

\[ = \hat{\zeta}_p + \frac{1}{f(\zeta_p)}(\hat{p} - p) \]

where the last equality follows from \( d\zeta_p/dp = (dp/d\zeta_p)^{-1} \) with \( p = F(\zeta_p) \). Therefore, we have

\[ \hat{\zeta}_p \approx \zeta_p + \frac{1}{f(\zeta_p)}(p - \hat{p}) \]  \hspace{1cm} (1)

which is first established by Bahadur (1966).
From (1), we can use

\[ V(\hat{\zeta}_p) \approx \{f(\zeta_p)\}^{-2} V(\hat{\rho}), \]

which requires knowledge of density function \( f(\cdot) \), or we need to estimate it.

Define \( Q(p) = F^{-1}(p) \). Since

\[ \frac{1}{f(\zeta_p)} = \frac{dQ(p)}{dp} \approx \frac{Q(p + h) - Q(p - h)}{p + h - (p - h)}, \]

we can use

\[ \frac{Q \left( \hat{\rho} + 2\sqrt{V(\hat{\rho})} \right) - Q \left( \hat{\rho} - 2\sqrt{V(\hat{\rho})} \right)}{\hat{\rho} + 2\sqrt{V(\hat{\rho})} - \{\hat{\rho} - 2\sqrt{V(\hat{\rho})}\}} := \hat{\gamma}_p \]

to estimate \( 1/f(\zeta_p) \). This is essentially the idea behind the test inversion method of Woodruff.
Bahadur Representation

\[ \sqrt{n}(\hat{\xi}_p - \xi_p) \to N \left( 0, \frac{p(1-p)}{[F'(\xi_p)]^2} \right) \] (SRS)

\[ V[F(\hat{\xi}_b)] = \left( \frac{\partial F}{\partial \xi} \right)^2 V(\hat{\xi}_p) \]

\[ \hat{Q}'(p) = [\hat{F}'(\xi_p)]^{-1} = \frac{\hat{Q}(\hat{\rho} + 2\sqrt{V(\hat{\rho})}) - \hat{Q}(\hat{\rho} - 2\sqrt{V(\hat{\rho})})}{(\hat{\rho} + 2\sqrt{V(\hat{\rho})}) - (\hat{\rho} - 2\sqrt{V(\hat{\rho})})} =: \hat{\gamma} \]

\[ \hat{V}(\hat{\xi}_p) = \hat{\gamma}^2 \hat{V}\{\hat{F}(\xi_p)\} \]
1. Jackknife variance estimator is not consistent. Median $\hat{\theta} = 0.5(x_m + x_{m+1})$ for $n=2m$ even

$$\hat{V}_{JK} = 0.25(n - 1)[x_{m+1} - x_m]^2$$

2. Bootstrap and BRR are O.K.

3. Smoothed quantile

$$\hat{\xi}_p = \hat{\xi}_{p0} + \hat{\gamma}[p - \hat{F}(\hat{\xi}_{p0})]$$

$$\hat{\xi}^{(k)}_p = \hat{\xi}_{p0} + \hat{\gamma}[p - \hat{F}^{(k)}(\hat{\xi}_{p0})]$$

$$\hat{V}_{JK}(\hat{\xi}_p) = \sum_k c_k(\hat{\xi}^{(k)}_p - \hat{\xi}_p)^2$$
Two-Phase Samples

\[ \bar{y}_{2p, \text{reg}} = \bar{y}_{2p, \text{REE}} = \bar{y}_2 + (\bar{x}_{1\pi} - \bar{x}_{2\pi})\hat{\beta}_2 \]

\[ \hat{\beta}_2 = \left\{ \sum_{i \in A_2} w_{2i} (x_i - \bar{x}_{2\pi})' (x_i - \bar{x}_{2\pi}) \right\}^{-1} \sum_{i \in A_2} w_{2i} (x_i - \bar{x}_{2\pi})' y_i \]

\[ w_{2i}^{-1} = \pi_{1i} \pi_{2i|1i}, \quad w_{1i} = \pi_{1i}^{-1} \]

\[ \hat{V}_{JK}(\bar{y}_{2p, \text{REE}}) = \sum_{k=1}^{L} c_k \left[ \bar{y}^{(k)}_{2p, \text{REE}} - \bar{y}_{2p, \text{REE}} \right]^2 \]
where \( w_{2i}^{(k)} = (w_{1i}^{(k)})^{(\pi_{2i}\mid 1i)} \)

\[
\bar{x}_{1\pi}^{(k)} = \left( \sum_{i \in A_1} w_{1i}^{(k)} \right)^{-1} \left( \sum_{i \in A_1} w_{1i}^{(k)} x_i \right)
\]

\[
(\bar{x}_{2\pi}^{(k)}, \bar{y}_{2\pi}^{(k)}) = \left( \sum_{i \in A_2} w_{2i}^{(k)} \right)^{-1} \left( \sum_{i \in A_2} w_{2i}^{(k)} (x_i, y_i) \right)
\]

\[
\hat{\beta}_2^{(k)} = \left[ \sum_{i \in A_2} w_{2i}^{(k)} (x_i - \bar{x}_{2\pi}^{(k)})'(x_i - \bar{x}_{2\pi}^{(k)}) \right]^{-1} \sum_{i \in A_2} w_{2i}^{(k)} (x_i - \bar{x}_{2\pi}^{(k)})'(y_i - \bar{y}_{2\pi}^{(k)})
\]
Theorem 4.2.1 Kim, Navarro, Fuller (2006 JASA), Assumptions

(i) Second phase is stratified with $\pi_{2i|1i}$ constant within group

(ii) $K_L < Nn^{-1}\pi_{1i} < K_U$ for some $K_L & K_U$

(iii) $V\{\hat{T}_{1y}|F\} \leq K_M V\{\hat{T}_{1y,SRS}\}$ where $\hat{T}_{1y} = \sum_{i \in A_1} \pi_{1i}^{-1}y_i$

(iv) $nV\{\hat{T}_{1y}|F\} = \sum_{i=1}^{N} \sum_{j=1}^{N} \Omega_{ij}y_iy_j$ where $\sum_{i=1}^{N} |\Omega_{ij}| = O(N^{-1})$

(v) $E\left[\left(\frac{\hat{V}_1(\hat{T}_{1y})}{V(\hat{T}_{1y}|F)} - 1\right)^2 |F\right] = o(1)$

(vi) $E\{[c_k(\hat{T}_{1y}^{(k)} - \hat{T}_{1y})^2]^2|F\} < K_LL^{-2}[V(\hat{T}_{1y})]^2$
$$\Rightarrow \hat{V}\{\bar{y}_{2p,\text{reg}}\} = V(\bar{y}_{2p,\text{reg}}|\mathcal{F}) - \frac{1}{N^2} \sum_{i=1}^{N} \kappa_{2i}^{-1}(1 - \kappa_{2i})e_i^2 + o_p(n^{-1})$$

where

- $\kappa_{2i} = \pi_{2i|1i}$
- $e_i = y_i - \bar{y}_N - (x_i - \bar{x})\beta_N$
- $\beta_N = \left[\sum_{i=1}^{N} (x_i - \bar{x}_N)'(x_i - \bar{x}_N)\right]^{-1} \sum_{i=1}^{N} (x_i - \bar{x}_N)'y_i$. 
Extend definition of $a_i$:

$$a_i = \begin{cases} 
1 & \text{if } i \in A_2, \\
0 & \text{if } i \notin A_2,
\end{cases} \begin{cases} 
1 & \text{if } i \in A_1, \\
0 & \text{if } i \notin A_1.
\end{cases}$$

$a_i$ defined throughout the population.

Writing

$$\bar{y}_{2p,\text{reg}} = \bar{z}_{1\pi} + (\bar{x}_{1\pi} - \bar{u}_{1\pi})\hat{\beta}$$

where $(\bar{z}_{1\pi}, \bar{u}_{1\pi}) = (\sum_{i \in A_1} a_{i\pi}^{-1}a_i^{-1} \sum_{i \in A_1} \pi_{2i}^{-1}(z_i, u_i), z_i = a_i y_i, u_i = a_i x_i)$

: conditional on $a = (a_1, \cdots, a_N)$, $\bar{y}_{2p,\text{reg}}$ is a smooth function of $\bar{z}_{1\pi}, \bar{x}_{1\pi}, \bar{u}_{1\pi}, \hat{\beta}$

$$\hat{\beta} = \left[ \sum_{i \in A_1} \pi_{2i}^{-1}(u_i - \bar{u}_{1\pi})'(u_i - \bar{u}_{1\pi}) \right]^{-1} \sum_{i \in A_1} \pi_{2i}^{-1}(u_i - \bar{u}_{1\pi})'z_i.$$
Variance of $\bar{y}_{2p,reg}$:

$$V[\bar{y}_{2p,reg}|\mathcal{F}] = E[V(\bar{y}_{2p,reg}|a, \mathcal{F})|\mathcal{F}] + V[E(\bar{y}_{2p,reg}|a, \mathcal{F})|\mathcal{F}]$$

Replication variance estimation:

$$\bar{y}_{2p,reg}^{(k)} = \bar{z}_{1\pi}^{(k)} + (\bar{x}_{1\pi}^{(k)} - \bar{u}_{1\pi}^{(k)})\hat{\beta}^{(k)}$$

$$\sum_{k=1}^{L} c_k (\bar{y}_{2p,reg}^{(k)} - \bar{y}_{2p,reg})^2 = E[V(\bar{y}_{2p,reg}|a, \mathcal{F})|\mathcal{F}] + o_p(n^{-1})$$

$$\hat{V}(\bar{y}_{reg}) = \sum_{k=1}^{L} c_k (\bar{y}_{2p,reg}^{(k)} - \bar{y}_{2p,reg})^2 =: \hat{V}(\bar{y}_{2p,reg}|a),$$

where $\hat{V}(\bar{y}_{2p,reg}|a)$ is used to emphasize that it is a function of $a$. 
Sketched Proof (Cont’d)

1. By condition (v), \( \hat{V}(\bar{y}_{2p, \text{reg}} | \mathbf{a}) = V[\bar{y}_{2p, \text{reg}} | \mathbf{a}, \mathcal{F}] + o_p(n^{-1}) \) conditional on \( \mathbf{a}, \mathcal{F} \).

2. Show \( \hat{V}(\bar{y}_{2p, \text{reg}} | \mathbf{a}) = E\{V[\bar{y}_{2p, \text{reg}} | \mathbf{a}, \mathcal{F}] | \mathcal{F}\} + o_p(n^{-1}) \)

   Need to show that \( \frac{\hat{V}(\bar{y}_{2p, \text{reg}} | \mathbf{a})}{V(\bar{y}_{2p, \text{reg}} | \mathbf{a}, \mathcal{F})} = o(1) \)

3. \[
\begin{align*}
\hat{V}(\bar{y}_{2p, \text{reg}} | \mathbf{a}) &= V(\bar{y}_{2p, \text{reg}} | \mathcal{F}) - V(E(\bar{y}_{2p, \text{reg}} | \mathbf{a}, \mathcal{F}) | \mathcal{F}) + o_p(n^{-1}) \\
\bar{y}_{2p, \text{reg}} &= \bar{y}_{2\pi} + (\bar{x}_{1\pi} - \bar{x}_{2\pi}) \hat{\beta} \\
&= \bar{y}_{2\pi} + (\bar{x}_{1\pi} - \bar{x}_{2\pi}) \beta_N + O_p(n^{-1}) \\
&= \bar{e}_{2\pi} + \bar{x}_{1\pi} \beta_N + O_p(n^{-1}) \\
E[\bar{y}_{2p, \text{reg}} | \mathbf{a}, \mathcal{F}] &= \frac{1}{N} \sum_{i=1}^{N} \kappa_{2i}^{-1} a_i e_i + \bar{x}_{1\pi} \beta_N \\
V[E[\bar{y}_{2p, \text{reg}} | \mathbf{a}, \mathcal{F}] | \mathcal{F}] &= \frac{1}{N^2} \sum_{i=1}^{N} \kappa_{2i}^{-1} (1 - \kappa_{2i}) e_i^2
\end{align*}
\]
Remark

(i) The bias term of $\hat{V}\{\bar{y}_{2p,\text{reg}}\}$ is of order $O(N^{-1})$ and can be negligible if $n/N$ is negligible.

(ii) The bias term can be unbiasedly estimated by

$$N^{-2} \sum_{i \in A_2} w_{1i} \pi_{2i|1i}^{-2} (1 - \pi_{2i|1i}) \hat{e}_i^2.$$

(iii) For some designs and replication procedures, it is possible to modify the weights to reduce the bias. (See the textbook.)