# Boundary Bias Correction Using the Weighting Method under Random Non Response in Two Stage Cluster Sampling with Replacement

Nelson K. Bii

Institute of Mathematical Sciences, Strathmore University

October 26, 2018

Supervisors: Dr. Christopher Ouma Onyango, Prof. John Odhiambo

# Introduction: Boundary Effects



- Let $m(x)$ denote a density function with support $[0, \infty)$ and consider non-parametric estimator of $m(x)$ based on a random sample $X_{ij}$. Define a standard kernel estimator of $m(x)$ by

$$\hat{m}(x) = \frac{1}{mnb} \sum_{i=1}^{n} \sum_{j=1}^{m} K\left(\frac{x_{ij} - X_{ij}}{b}\right). \quad (1)$$

- $\hat{m}(x)$ underestimates $m(x)$ on the set $[0, b)$ since the kernel estimator $\hat{m}(x)$ does not detect data outside the boundaries of the support and so naturally it penalizes the density estimate within the $b$-neighborhood of zero.
- Hence at the boundary region ($x \in [0, b)$), the estimator is inconsistent and has a boundary bias; these are called boundary effects.

**Reflection of Data Method**

- Since the kernel estimator is penalizing for lack of data on the negative axis, just add $-X_1, -X_2, \cdots, -X_n$ to the data set. The reflections result in a new data set $(-X_1, X_1), (-X_2, X_2), \cdots, (-X_n, X_n)$, for details see [5], [6] and [7].
- Estimator of $m(x)$ becomes

$$\hat{m}(x) = \frac{1}{nb}\sum_{i=1}^{n}\left\{K\left(\frac{x - X_i}{b}\right) + K\left(\frac{x + X_i}{b}\right)\right\}. \qquad (2)$$

for $x \geq 0$, $\hat{m}\prime(0) = 0$

- Method recommended if the property $\hat{m}\prime(0) = 0$ holds.

# Methods of reducing boundary Effects...



**Transformation of Data Method**

- Take a one-to-one continuous function $g$ on $[0, \infty) \to [0, \infty)$
- Transform original data $X_1, X_2, \cdots, X_n$ to $g(X_1), g(X_2), \cdots, g(X_n)$
- Use the regular kernel estimator with the transformed data set $g(X_1), g(X_2), \cdots, g(X_n)$.

# Transformation of Data Method...



- Estimator of $m(x)$ becomes

$$\hat{m}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x_i - g(X_i)}{b}\right). \tag{3}$$

- For more details see [4] and [3].
- Produces non-negative estimates with low variance.

# Proposed Method



**Modified Transformation of Data Method**

- Generate data beyond the left end point of the support of the density function.
- Transform original data $X_{ij}$ to $g(X_{ij})$.
- Reflect $g(X_{ij})$ around the origin so that we have $(-g(X_{ij}), g(X_{ij}))$. Estimator of $m(x)$ becomes

$$m_{TDM}(\hat{x_{ij}}) = \frac{1}{mnb}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{K\left(\frac{x_{ij}-X_{ij}}{b}\right)+K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right\} \quad (4)$$

# Estimation of Finite Population Mean, $\overline{\overline{Y}}$



- Estimation of finite population mean using kernel density estimators in presence of non-response lead to bias due to boundary effects, see for instance [1].
- Data is generated using a regression model applied by [2]

$$Y_{ij} = m(x_{ij}) + e_{ij}. \tag{5}$$

- $m(x_{ij})$ is estimated by (4).
- Estimator of $\overline{\overline{Y}}$ is given by

$$\hat{\overline{\overline{Y}}} = \frac{1}{M}\left\{ \sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} + \sum_{j=n+1}^{N}\sum_{i=m+1}^{M} \hat{y}_{ij} \right\}. \tag{6}$$

# Estimation of finite population mean, $\overline{\overline{Y}}$



- where

$$\hat{y}_{ij} = \sum_i \sum_j W^*_{ij}(x_{ij}) Y_{ij}. \tag{7}$$

and

$$W^*_{ij}(x_{ij}) = \frac{\left\{ K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right) \right\}}{\sum_{i=n+1}^{N} \sum_{j=m+1}^{M} \left\{ K\left(\frac{x_{ij}-X_{ij}}{b}\right) + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right) \right\}}. \tag{8}$$

- Hence estimator of $\overline{\overline{Y}}$ is given by

$$\hat{\overline{\overline{Y}}} = \frac{1}{Mn} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij} + \sum_{j=n+1}^{N} \sum_{i=m+1}^{M} W^*_{ij}(x_{ij}) Y_{ij} \right\} \tag{9}$$

# Asymptotic Bias



$$E\left(m_{TDM}(\hat{x_{ij}})\right) - m(x_{ij}) = b\int_a^1 \left(w-a\right)K(w)dw\Big[2m\prime(0) - g\prime\prime(0)m(0)\Big] + \frac{b^2}{2}m\prime\prime(0)\int_{-1}^1 w^2K(w)dw - \frac{b^2}{2}\int_a^1 \left(w-a\right)^2 K(w)dw\Big\{g\prime\prime(0)m(0) + 3g\prime\prime(0)\Big[m\prime(0) - g\prime\prime(0)m(0)\Big]\Big\} + o(b^2). \tag{10}$$

- As $b \to 0$, it is noted that $E\left(m_{TDM}(\hat{x_{ij}})\right) - m(x_{ij}) \approx 0$.

# Asymptotic Variance



$$Var\left(m_{TDM}(\hat{x_{ij}})\right) = \frac{1}{(mn)^2 b^2} \sum_{i=n+1}^{N} \sum_{j=m+1}^{M} E\Bigg\{ K\left(\frac{x_{ij} - X_{ij}}{b}\right) + K\left(\frac{x_{ij} + g(X_{ij})}{b}\right) - E\left[K\left(\frac{x_{ij} - X_{ij}}{b}\right) + K\left(\frac{x_{ij} + g(X_{ij})}{b}\right)\right]\Bigg\}^2 . \tag{11}$$

# Variance...



- Equation (11) can be re-written as

$$
\begin{aligned}
Var\left(m_{TDM}(\hat{x_{ij}})\right) = & \frac{(Mn-mn)^2}{(mn)^2 b^2} E\Bigg[K\left(\frac{x_{ij}-X_{ij}}{b}\right) \\
& + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\Bigg]^2 \\
& - \frac{(Mn-mn)^2}{(mn)^2 b^2}\Bigg\{E\Bigg[K\left(\frac{x_{ij}-X_{ij}}{b}\right) \\
& + K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\Bigg]\Bigg\}^2 \\
= & A+B.
\end{aligned} \tag{12}
$$

# Variance...

$$
\begin{aligned}
A &= \frac{(Mn - mn)^2}{mnb^2} \int \left[ K\left(\frac{x_{ij} - y_{ij}}{b}\right) + K\left(\frac{x_{ij} + g(y_{ij})}{b}\right) \right]^2 f(y)dy \\
&= \frac{(Mn - mn)^2}{mnb^2} \left\{ \int \left[ K\left(\frac{x_{ij} - y_{ij}}{b}\right)^2 f(y)dy + \int K\left(\frac{x_{ij} + g(y_{ij})}{b}\right) \right. \right. \\
&\quad \left. \left. \times f(y)dy \right] + 2 \int K\left(\frac{x_{ij} - y_{ij}}{b}\right) K\left(\frac{x_{ij} + g(y_{ij})}{b}\right) f(y)dy \right\}. \qquad (13)
\end{aligned}
$$

# Variance...

$$B = -\frac{(Mn-mn)^2}{(mn)^2 b^2}\left\{E\left[K\left(\frac{x_{ij}-X_{ij}}{b}\right)+K\left(\frac{x_{ij}+g(X_{ij})}{b}\right)\right]\right\}^2. \tag{14}$$

- Using Taylor's series expansions, $A$ and $B$ can be expanded and simplified to obtain

$$\begin{aligned} Var\left(m_{TDM}(\hat{x_{ij}})\right) &= A + B \\ &= \frac{\left(Mn-mn\right)^2}{mnb}\left[\int_{-1}^{1} K(w)^2 dw \right. \\ &\left. + 2\int_{-1}^{a} K(w)K\left(2a-w\right)dw\right] + O\left(\frac{1}{mnb}\right). \end{aligned} \tag{15}$$

# Conclusion



- It is noted that the the variance is decreasing in $mnb$ whereas the bias is increasing in $b$. Both the bias and the variance must become small as $mn \to \infty$ for the estimator to be optimal.
- That is for all $x_{ij}$, $m_{TDM}(\hat{x_{ij}}) \to m(x_{ij})$ in probability as $mn \to \infty$.

# References



[1] Bii, N. K., Onyango, C. O., & Odhiambo, J. (2017). *Estimating a Finite Population Mean under Random Non-Response in Two Stage Cluster Sampling with Replacement*. Open Journal of Statistics, 7(05), 834.

[2] Onyango, C. O., Otieno, R. O., & Orwa, G. O. (2010). *Generalised Model Based Confidence Intervals in Two Stage Cluster Sampling*. Pakistan Journal of Statistics and Operation Research, 6(2), 101-115.

[3] Marron, J. S., & Ruppert, D. (1994). *Transformations to reduce boundary bias in kernel density estimation*. Journal of the Royal Statistical Society. Series B (Methodological), 653-671.

[4] Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

# References

[5] Silverman, B. W. (2018). *Density estimation for statistics and data analysis. Routledge*.

[6] Simonoff, J. S. (2012). Smoothing methods in statistics. Springer Science & Business Media.

[7] Ivanka, H., Jan, K., & Jiri, Z. (2012). Kernel Smoothing in MATLAB: theory and practice of kernel smoothing. World scientific.