Weight Smoothing via Design Modeling in Complex Surveys

Jay Breidt

NORC at the University of Chicago

Joint Statistical Meetings

August 5, 2024

thanks to Ben Reist and Taylor Wing (NORC) and many NASS collaborators

Stratified finite population setting

- Finite population $U = \{1, 2 \ldots, N\}$
- Conditioning on $\boldsymbol{y}$, with nonrandom inferential target:

$$T_y = \sum_{k \in U} y_k = \sum_{k \in U} \sum_{j=1}^{J} y_k L_{jk}$$

  - disjoint stratum indicators L_{jk} with $\sum_{j=1}^{J} L_{jk} = 1$
- Independent samples $\{I_{jk}\}$ across strata
- $\Pr[I_{jk} = 1 \mid L_{jk} = 1] = \pi_{jk} > 0$, not random (not SOIP!)
- Traditional Horvitz-Thompson unbiased estimator:

$$\widehat{T}_y = \sum_{k \in U} \sum_{j=1}^{J} y_k \frac{I_{jk} L_{jk}}{\pi_{jk}}$$

Stratified finite population setting, continued

- Now suppose that strata are not necessarily predictive of $\boldsymbol{y}$-variation . . .
  - J deduplicated frames that together cover U completely
  - screening strata for a domain of interest, with range of anticipated hit rates
  - mis-specified strata; "stratum jumpers"
- . . . and the weights vary considerably across strata:

$$\sum_{j=1}^{J} \frac{I_{jk} L_{jk}}{\pi_{jk}}$$

- This leads to inefficiency of traditional HT estimator

Smoothed Horvitz-Thompson estimator

- Beaumont (2008, *Biometrika*) proposed **smoothed HT**:

$$\mathrm{E}\left[\sum_{k\in U}\pi_k^{-1}y_k I_k \,\middle|\, s, \boldsymbol{y}\right] = \sum_{k\in U}\mathrm{E}\left[\pi_k^{-1}\middle|\, s, \boldsymbol{y}\right] y_k I_k$$

  - smoothed weight $\mathrm{E}\left[\pi_k^{-1}\middle|\, s, \boldsymbol{y}\right]$ averages out variation not predictive of $\boldsymbol{y}$
  - **Rao-Blackwell:** if smoothed weights were known, smoothed HT is unbiased and has variance no larger than HT
  - introduce parametric models for weights and smooth them accordingly (within-sample modeling)

What does conditioning on s mean in our context?

- Compute conditional expectation (always conditional on $\boldsymbol{y}$):

$$\mathrm{E}\left[\widehat{T}_y \,\middle|\, s\right] = \sum_{k \in U} \mathrm{E}\left[\sum_{j=1}^{J} \frac{I_{jk} L_{jk}}{\pi_{jk}} y_k \,\middle|\, s\right]$$

- Hence the smoothed weight is

$$\mathrm{E}\left[\sum_{j=1}^{J} \frac{I_{jk} L_{jk}}{\pi_{jk}} \,\middle|\, s\right]$$

- Because the strata are not necessarily predictive of $\boldsymbol{y}$-variation, we want to average them out: we only care that

$$I_k = \sum_{j=1}^{J} I_{jk} L_{jk} = 1,$$

not *which stratum* brought element k into the sample

Smoothed HT in our context, continued

- Accordingly, condition on $I_k = 1$ (and implicitly on $\boldsymbol{y}$):

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{J} \frac{I_{jk} L_{jk}}{\pi_{jk}} \,\middle|\, I_k = 1\right] &= \sum_{j=1}^{J} \frac{1}{\pi_{jk}} \Pr\left[I_{jk} L_{jk} = 1 \mid I_k = 1\right] \\
&= \sum_{j=1}^{J} \frac{1}{\pi_{jk}} \frac{\Pr\left[I_k = 1 \mid I_{jk} L_{jk} = 1\right] \Pr\left[I_{jk} L_{jk} = 1\right]}{\Pr\left[I_k = 1\right]} \\
&= \sum_{j=1}^{J} \frac{1}{\pi_{jk}} \frac{1 \cdot \Pr\left[I_{jk} = 1 \mid L_{jk} = 1\right] \Pr\left[L_{jk} = 1\right]}{\Pr\left[I_k = 1\right]} \\
&= \sum_{j=1}^{J} \frac{1}{\pi_{jk}} \frac{\pi_{jk} \lambda_{jk}}{\Pr\left[I_k = 1\right]} = \frac{1}{\sum_{j=1}^{J} \pi_{jk} \lambda_{jk}} = \frac{1}{\pi_{\bullet k}}
\end{aligned}$$

Multinomial "design model" of stratum membership

- However k enters the sample, its smoothed weight is

$$\frac{1}{\Pr[I_k = 1]} = \frac{1}{\Pr[k \in s]} = \frac{1}{\sum_{j=1}^{J} \pi_{jk}\lambda_{jk}} = \frac{1}{\pi_{\bullet k}}$$

- Unlike Beaumont (2008), no sample-level model for π_{jk}
- Instead, population-level **design model** for stratum membership:

$$(L_{1k}, L_{2k}, \ldots, L_{Jk}) \sim \text{independent Multinomial}(\boldsymbol{\lambda}_k)$$

  - stratum probabilities $(\lambda_{1k}, \lambda_{2k}, \ldots, \lambda_{Jk})$ can depend on any $\boldsymbol{y}$ observed in the sample
  - introduce parametric model for $\boldsymbol{\lambda}_k$ and estimate via sample-weighted likelihood

Some theory for the smoothed HT estimator: bias

- Estimator $\widetilde{T}_y = \sum_{k \in U} y_k I_k \pi_{\bullet k}^{-1}$
- Design biased for T_y, conditional on $\boldsymbol{L}$:

$$\mathrm{E}\left[\left(\widetilde{T}_y - T_y\right) \middle| \boldsymbol{L}\right] = \sum_{j=1}^{J} \sum_{k \in U} y_k \left(\frac{\pi_{jk}}{\pi_{\bullet k}} - 1\right) L_{jk} \neq 0$$

- Design bias has mean zero, averaging over $\boldsymbol{L}$:

$$\mathrm{E}\left[\mathrm{E}\left[\left(\widetilde{T}_y - T_y\right) \middle| \boldsymbol{L}\right]\right] = 0$$

- Design bias converges in mean square to zero

$$\mathrm{E}\left[\mathrm{E}\left[N^{-2}\left(\widetilde{T}_y - T_y\right)^2 \middle| \boldsymbol{L}\right]\right] \to 0 \text{ as } N \to \infty$$

  under standard design asymptotics

Some theory for the smoothed HT estimator: variance

- Design variance is conditional on $\boldsymbol{L}$:

$$\text{Var}\left(\widetilde{T}_y \mid \boldsymbol{L}\right) = \sum_{j=1}^{J} \sum_{k,\ell \in U} \Delta_{j,k\ell} \frac{y_k}{\pi_{\bullet k}} \frac{y_\ell}{\pi_{\bullet \ell}} L_{jk} L_{j\ell}$$

  - just like stratified sampling, but with $\pi_{\bullet k}$ replacing π_{jk}
  - easy to adapt design-based methods to estimate $\text{Var}\left(\widetilde{T}_y \mid \boldsymbol{L}\right)$
  - unbiased estimator of $\text{Var}\left(\widetilde{T}_y \mid \boldsymbol{L}\right)$ is also unbiased for $\text{E}\left[\text{Var}\left(\widetilde{T}_y \mid \boldsymbol{L}\right)\right]$

- Also straightforward to obtain consistent estimator of

$$\text{Var}\left(\text{E}\left[\widetilde{T}_y \mid \boldsymbol{L}\right]\right) = \sum_{k \in U} \left(\frac{y_k}{\pi_{\bullet k}}\right)^2 \text{Var}\left(\sum_{j=1}^{J} \pi_{jk} L_{jk}\right)$$

Special case: stratified sampling for domains

- λ_{jk} can depend on any $\boldsymbol{y}$ observed in the sample
- Suppose λ_{jk} depends only on **domain** $(D_{1k}, D_{2k}, \ldots, D_{Ik})$
- By Bayes' rule,

$$\lambda_{jk} = \Pr\left[L_{jk} = 1 \mid D_{ik} = 1\right] = \frac{\Pr\left[D_{ik} = 1 \mid L_{jk} = 1\right] \Pr\left[L_{jk} = 1\right]}{\Pr\left[D_{ik} = 1\right]},$$

- Whatever the original stratum, same smoothed domain weight:

$$\frac{1}{\sum_{j=1}^{J} \pi_{jk} \lambda_{jk}} = \frac{\Pr\left[D_{ik} = 1\right]}{\sum_{j=1}^{J} \pi_{jk} \Pr\left[D_{ik} = 1 \mid L_{jk} = 1\right] \Pr\left[L_{jk} = 1\right]}$$

$$= \frac{(\text{population frequency of domain } i)}{\sum_{j=1}^{J} \pi_{jk} \, (\text{hit rate of } i \text{ in stratum } j) \, (\text{frame proportion of } j)}$$

Stratified sampling for domains: useless strata

- Stratify and look for domains D1, D2, D3

			Domain			
			D1	D2	D3	Not
			0.13	0.09	0.33	0.45
Stratum		**Weight**				
Likely D1	0.10	**5.3**	0.13	0.09	0.33	0.45
Likely D2	0.07	**2.9**	0.13	0.09	0.33	0.45
Likely D3	0.35	**4.7**	0.13	0.09	0.33	0.45
Likely Not	0.48	**15.1**	0.13	0.09	0.33	0.45

Stratified sampling for domains: useless strata

- Domain weights are **completely smoothed**

			Domain			
			D1	D2	D3	Not
			0.13	0.09	0.33	0.45
Stratum		**Smooth**	**6.67**	**6.67**	**6.67**	**6.67**
Likely D1	0.10	**5.3**	0.13	0.09	0.33	0.45
Likely D2	0.07	**2.9**	0.13	0.09	0.33	0.45
Likely D3	0.35	**4.7**	0.13	0.09	0.33	0.45
Likely Not	0.48	**15.1**	0.13	0.09	0.33	0.45

Stratified sampling for domains: imperfect strata

- Stratify and look for domains D1, D2, D3

			Domain			
			D1	D2	D3	Not
			0.13	0.09	0.33	0.45
Stratum		**Weight**				
Likely D1	0.10	**5.3**	0.64	0.03	0.08	0.25
Likely D2	0.07	**2.9**	0.03	0.56	0.11	0.30
Likely D3	0.35	**4.7**	0.06	0.05	0.68	0.21
Likely Not	0.48	**15.1**	0.09	0.06	0.15	0.70

Stratified sampling for domains: imperfect strata

- Domain weights are **partially smoothed**

			Domain			
			D1	D2	D3	Not
			0.13	0.09	0.33	0.45
Stratum		**Smooth**	**6.5**	**4.4**	**5.4**	**9.1**
Likely D1	0.10	**5.3**	0.64	0.03	0.08	0.25
Likely D2	0.07	**2.9**	0.03	0.56	0.11	0.30
Likely D3	0.35	**4.7**	0.06	0.05	0.68	0.21
Likely Not	0.48	**15.1**	0.09	0.06	0.15	0.70

Stratified sampling for domains: perfect strata

- Stratify and look for domains D1, D2, D3

			Domain			
			D1	D2	D3	Not
			0.10	0.07	0.35	0.48
Stratum		**Weight**				
Likely D1	0.10	**5.3**	1	0	0	0
Likely D2	0.07	**2.9**	0	1	0	0
Likely D3	0.35	**4.7**	0	0	1	0
Likely Not	0.48	**15.1**	0	0	0	1

Stratified sampling for domains: perfect strata

- Domain weights are **completely unsmoothed**

			Domain			
			D1	D2	D3	Not
			0.10	0.07	0.35	0.48
Stratum		**Smooth**	**5.3**	**2.9**	**4.7**	**15.1**
Likely D1	0.10	**5.3**	1	0	0	0
Likely D2	0.07	**2.9**	0	1	0	0
Likely D3	0.35	**4.7**	0	0	1	0
Likely Not	0.48	**15.1**	0	0	0	1

National Agricultural Statistics Service applications

- Some of the most important NASS farm surveys use both
  - **list frame** with reasonable coverage of farms
  - **area frame** to address undercoverage
- Dual-frame surveys cover range of agricultural topics:
  - Livestock: Cattle, Hogs, Sheep and Goats
  - Labor, ARMS (Agricultural Resource Management Survey)
  - Crop APS (Acreage, Production, and Stocks)
- Area frame sample is **expensive**
  - stratify the landscape
  - select stratified simple random sample of area segments
  - map selected area segments, finding agricultural **tracts**
  - determine if tract corresponds to a listed farm or NOL (not-on-list)

Farm example, I

- A simple example with invented sampling probabilities:
  - list sample selects large farms with probability $\pi_{1k} = 0.5$
  - list sample selects small farms with probability $\pi_{1k} = 0.2$
  - area sample selects all NOL farms with probability $\pi_{2k} = 0.01$
- The corresponding weights are then as follows:

Estimator	Farm Size	Frame	Weight
NASS	Large	List	$1/0.5 = 2$
		Area	$1/0.01 = 100$
	Small	List	$1/0.2 = 5$
		Area	$1/0.01 = 100$

- An (unlisted) area-frame large farm can make a huge contribution to the estimate

Weight variation in NASS surveys

- Large NOL contributions to the estimates and their variability are common in NASS surveys
  - often fail to meet precision targets due to NOL
- NASS cattle surveys:

$$\begin{aligned}(\text{all states}) \quad &\times \quad (\text{two surveys per year}) \\ &\times \quad (\text{multiple characteristics per survey})\end{aligned}$$

- Addressed via weight trimming or other post-hoc methods
- Serious and time-consuming issue under tight production timelines

Farm example, II

- Now suppose we try to account for the fact that missing a large farm is not as likely as missing a small farm
- Same example with invented sampling probabilities:
  - list sample selects large farms with probability $\pi_{1k} = 0.5$
  - list sample selects small farms with probability $\pi_{1k} = 0.2$
  - area sample selects all NOL farms with probability $\pi_{2k} = 0.01$
- Add information on listing probabilities:
  - large farms are listed with probability $\lambda_{1k} = 0.965$ (very likely to be listed, but not perfect)
  - small farms are listed with probability $\lambda_{1k} = 0.838$
  - NOL probabilities are $\lambda_{2k} = 1 - \lambda_{1k}$

Farm example, III

- The smoothed HT weights are better-behaved:

Estimator	Farm Size	Frame	Weight
NASS (HT)	Large	List	$1/0.5 = 2$
		Area	$1/0.01 = 100$
	Small	List	$1/0.2 = 5$
		Area	$1/0.01 = 100$
Smoothed HT	Large	List or Area	$\frac{1}{(0.5)(0.965)+(0.01)(0.035)} = 2.07$
	Small	List or Area	$\frac{1}{(0.2)(0.838)+(0.01)(0.162)} = 5.91$

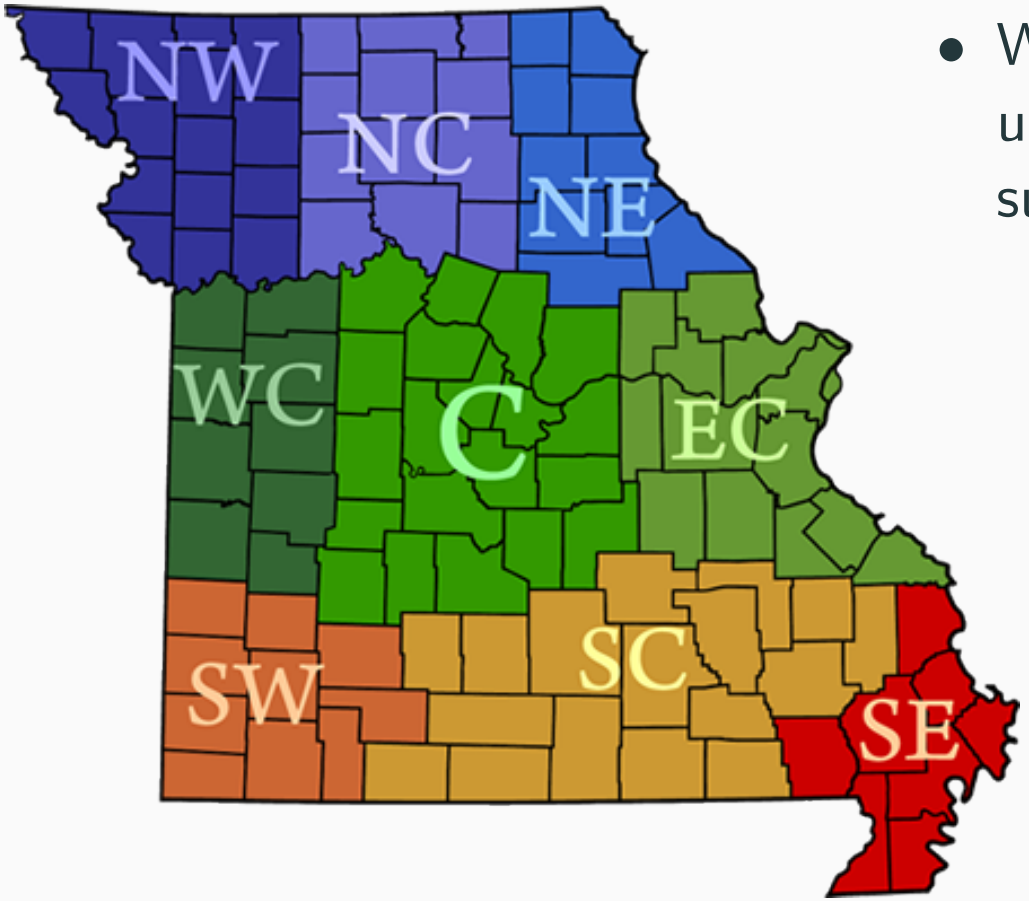
Implementation details

- **Smoothed HT estimator** of T_y:

$$\widetilde{T}_y = \sum_{k \in U} y_k \frac{I_{1k} L_{1k} + I_{2k} L_{2k}}{\pi_{1k} \lambda_{1k} + \pi_{2k}(1 - \lambda_{1k})} = \sum_{k \in U} y_k \frac{I_k}{\pi_{\bullet k}}$$

- Requires **coverage model** for listing probabilities λ_{1k}
  - (ok to depend on anything observable in combined sample)
- Requires list frame probabilities π_{1k} for NOL farms (as if they were listed)
- Requires NOL probabilities π_{2k} for list farms (as if they were NOL)
  - needs some modeling/approximation for both NOL and list

Modeling the listing probabilities for Missouri cattle



- Weighted logistic regression using two previous years of survey data
  - agricultural district
  - cattle variables: total, beef cows, milk cows, cattle on feed
  - livestock: hogs, horses
  - land acres: owned, rented from, rented to, operated
  - CRP (Conservation Reserve Program)
  - cropland acres

Monte Carlo experiment

- Simulate population of $N = 44,870$ **cattle farms**
  - generate $\boldsymbol{y}$ using scaled hot-deck donors from sample data as Poisson mean vectors
  - generate $\boldsymbol{L}$ using "true" coverage model (fitted to sample data)

Year	Data	Coverage	Sample(s)	Estimates
1	$\boldsymbol{y}_1$	$\boldsymbol{L}_1$	one sample, s_1	$\widehat{\lambda}^{(c)}, \widehat{\lambda}^{(ic1)}$
2	$\boldsymbol{y}_2$	$\boldsymbol{L}_2$	one sample, s_2	$\widehat{\lambda}^{(ic2)}, \widehat{\lambda}^{(ic3)}$
3	$\boldsymbol{y}$	$\boldsymbol{L}$	1000 MC samples, s (conditional on $\boldsymbol{y}, \boldsymbol{L}$)	$\widehat{T}_{\boldsymbol{y}}, \widetilde{T}_{\boldsymbol{y}}$

Bias properties

- HT is unbiased in simulation, modulo area weight approximation for NOL cases
- Smoothed HT is **biased** in simulation because Monte Carlo is conditional on $\boldsymbol{L}$
- Percent relative biases: $(100\%) \times (\text{MC mean} - \text{truth})/(\text{truth})$

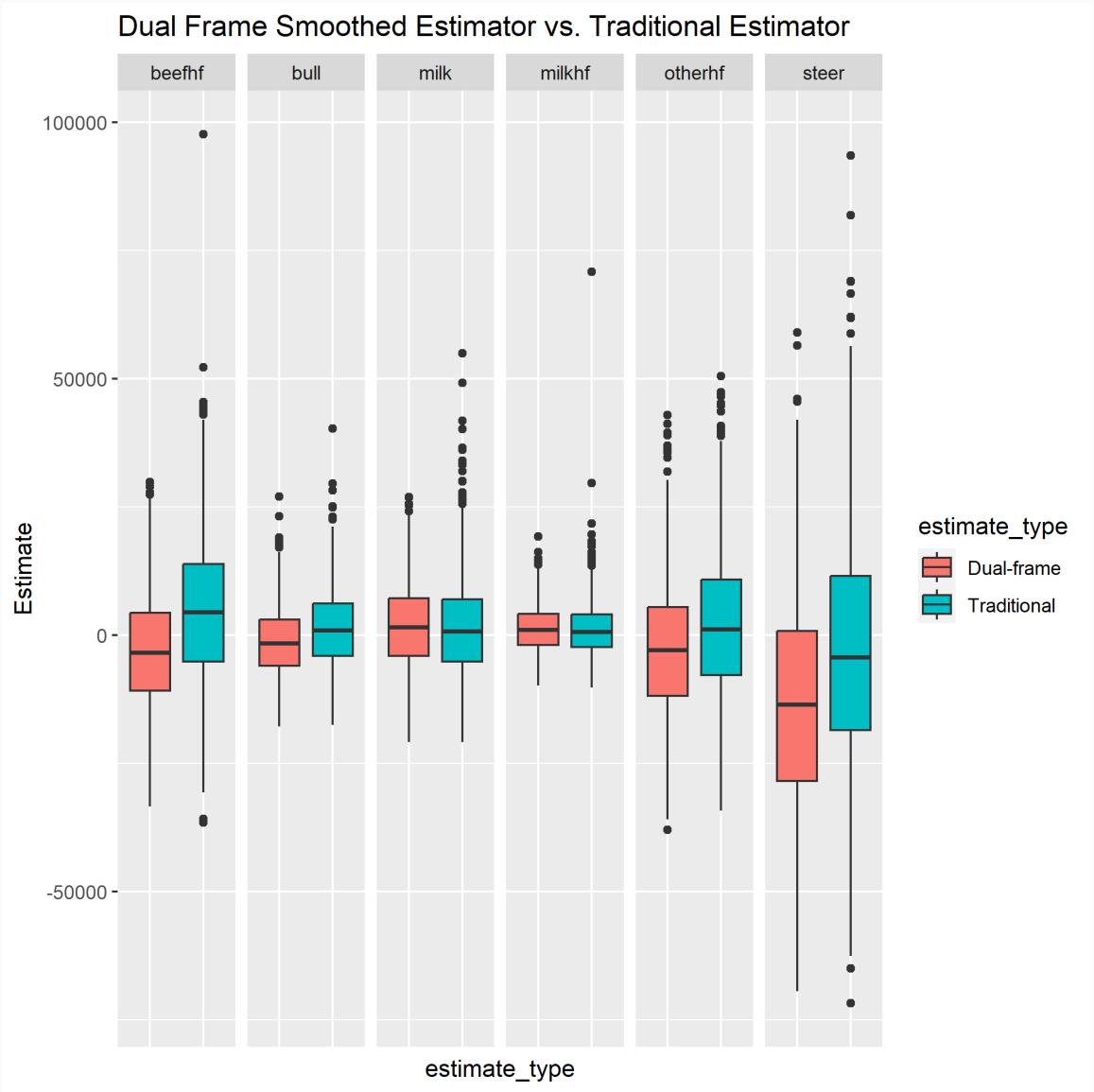
Drop	b.cow	m.cow	bull	b.hf	m.hf	o.hf	steer	calf
—	−1.9	2.4	−1.1	−1.3	4.3	−1.3	−3.1	−1.8
Dist	−2.2	2.4	−1.5	−1.5	4.2	−1.2	−3.1	−2.0
+CRP	−2.2	2.3	−1.5	−1.5	4.0	−1.0	−3.0	−2.0
+milk	−2.0	**7.4**	−1.3	−1.3	**8.1**	−0.9	−2.9	−1.7
NASS	0.8	2.0	1.3	2.1	4.2	0.9	−0.7	0.8

Root Mean Square Error properties

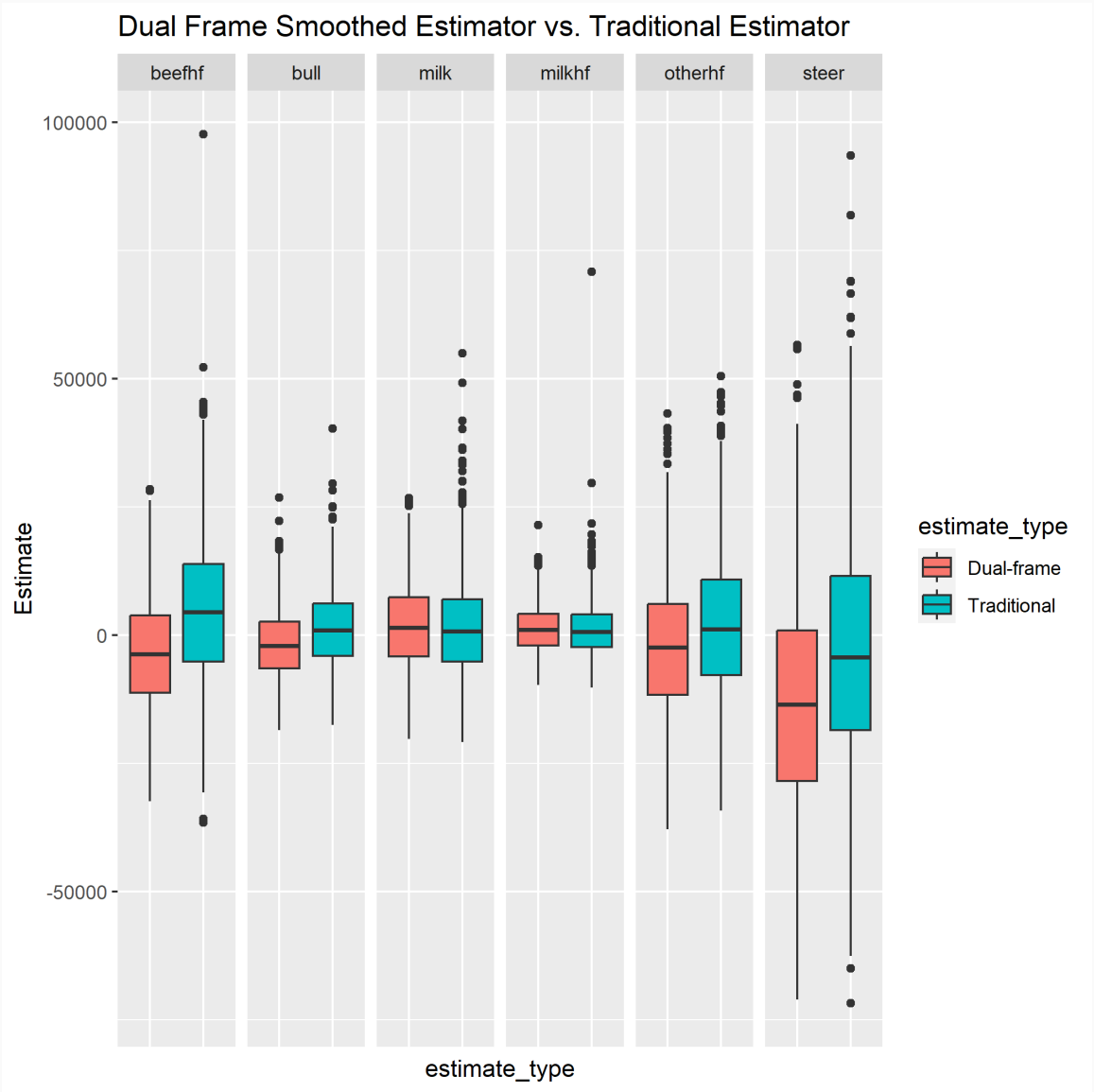
- RMSE ratios = (traditional)/(smoothed)

Drop	farm	b.cow	m.cow	bull	b.hf	m.hf	o.hf	steer	calf
—	2.10	1.15	1.17	1.14	1.30	1.20	1.05	0.91	1.26
Dist	1.62	1.10	1.16	1.13	1.31	1.21	1.06	0.92	1.22
+CRP	1.55	1.11	1.17	1.14	1.32	1.21	1.07	0.93	1.23
+milk	1.66	1.15	**0.96**	1.15	1.33	**1.09**	1.08	0.94	1.30

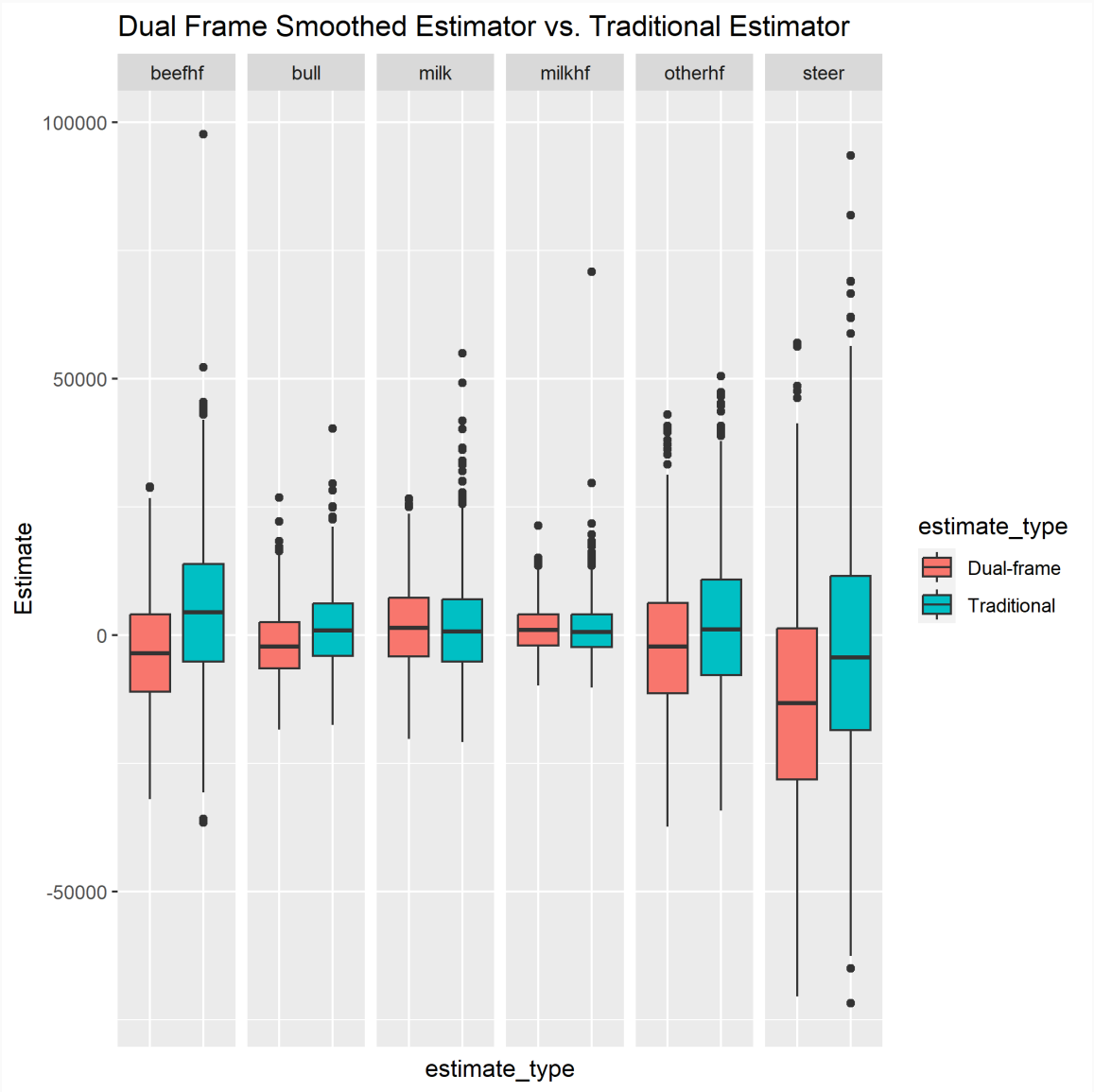
Correctly specified coverage model: $\widehat{\lambda}^{(c)}$



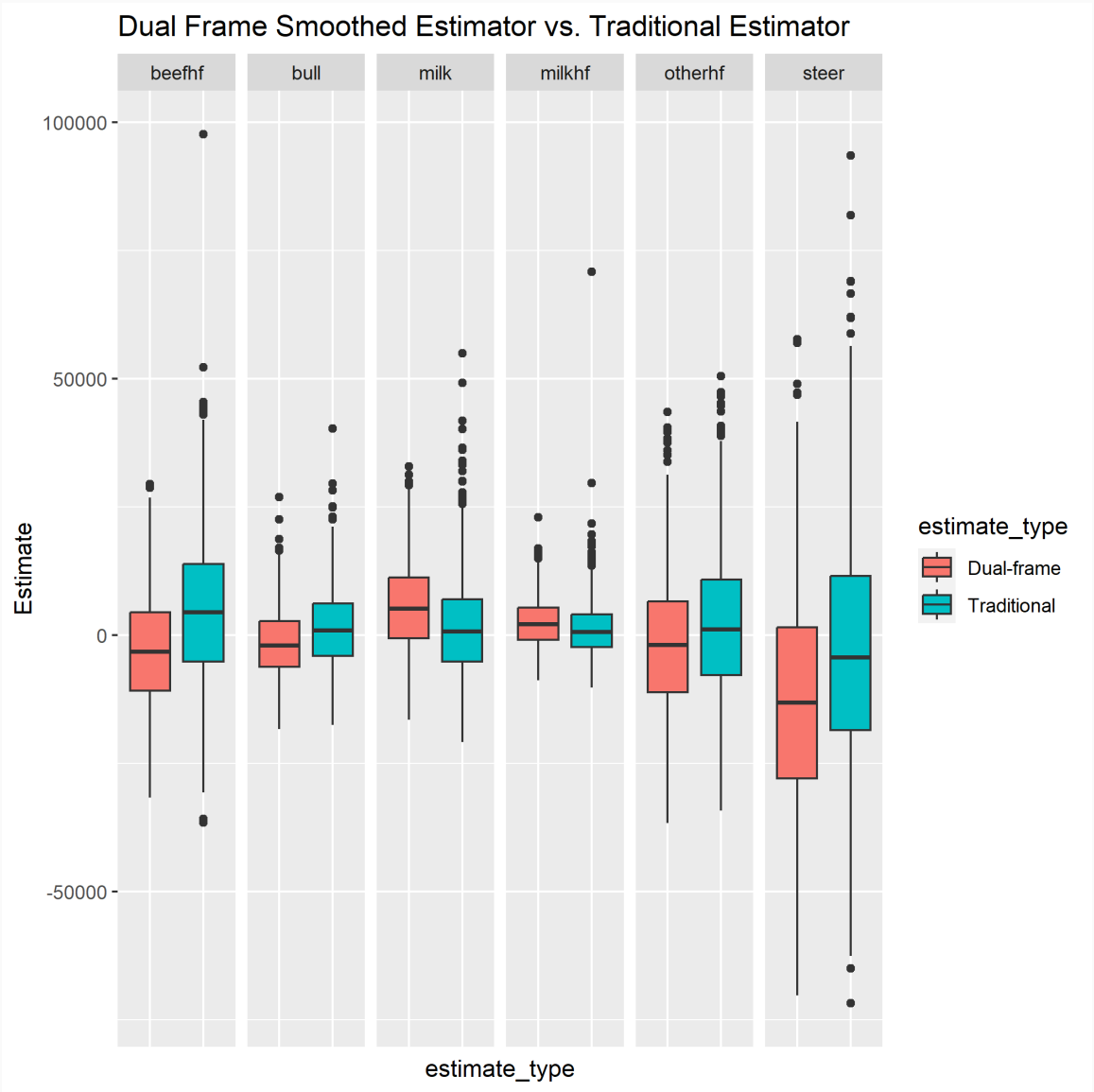
Dropping DISTRICT: $\widehat{\lambda}^{(ic1)}$



Dropping DISTRICT and CRP: $\widehat{\lambda}^{(ic2)}$



Dropping DISTRICT, CRP, and MILK: $\widehat{\lambda}^{(ic3)}$



Conclusion

- Weight smoothing is a principled approach to dealing with stratification uncertainty
- Closely tied to design-based paradigm, but uses population-level "design modeling"
  - design model can use information external to the survey
- In a realistic simulation motivated by NASS farm surveys:
  - reasonable bias properties, even with some model misspecification
  - better RMSE properties than traditional HT estimator used by NASS
  - far fewer outliers
  - NASS actively considering this methodology for surveys with NOL components