Precision of Estimates of Nonresponse Bias in Means

Stephanie Eckman

Jennifer Unangst, Jill Dever, Chris Antoun

RTI International is a registered trademark and a trade name of Research Triangle Institute.

www.rti.org

Motivation

- Published estimates of NR bias
  - Judge quality of that survey
  - Predict nonresponse patterns in future surveys

Example: repeated cross-sectional survey

Round X shows low response from HHs in Northwest region

$bias(p_r^{nw}) = p_r^{nw} - p_{full}^{nw}$

Should we change protocol in Round Y?

Motivation

- Maybe response is low because of *sample*
  - If we repeated survey, how would bias differ?

sampling variability of *NR bias estimates*

Research Questions

1. How variable are NR bias estimates?
   – RR
   – Clustering

2. How to estimate $Var(bias(\bar{y}_r))$?

Simulation Set Up: Unclustered

variable of interest
~ response propensity

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N\left[\begin{pmatrix} 10 \\ \delta \end{pmatrix}, \begin{pmatrix} \theta & \rho \\ \rho & 1 \end{pmatrix}\right]$$

Parameters			
$\boldsymbol{\delta}$	RR	*30%*	*70%*
$\boldsymbol{\rho}$	NR bias	*None*	*Up to 10% NR bias*

Simulation Set Up: Clustered

variable of interest
~ response propensity

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \sim N\left[\begin{pmatrix} 10+\phi \\ \delta \end{pmatrix}, \begin{pmatrix} \theta & \rho \\ \rho & 1 \end{pmatrix}\right]$$

Parameters			
δ	RR	*30%*	*70%*
ρ	NR bias	*None*	*Up to 10% NR bias*
φ	Clustering	*Unclustered*	*Clustered*

More Simulation Setup

- Used similar process to create binomial & Poisson Ys
  - Correlated with Z
  - Normal → uniform → binomial

$$Y \sim Bin(1, 0.5)$$
$$Y \sim Poisson(3)$$

- 2,000 samples of *n=1,000*
  - SRS
  - Multi-stage cluster

- Response propensity $RP_k = \frac{e^{1+Z_k}}{1+e^{1+Z_k}}$

  deterministic: $R_{det,k} = 0,1$

  stochastic: $R_{sto,s,k} = 0,1$

Research Questions

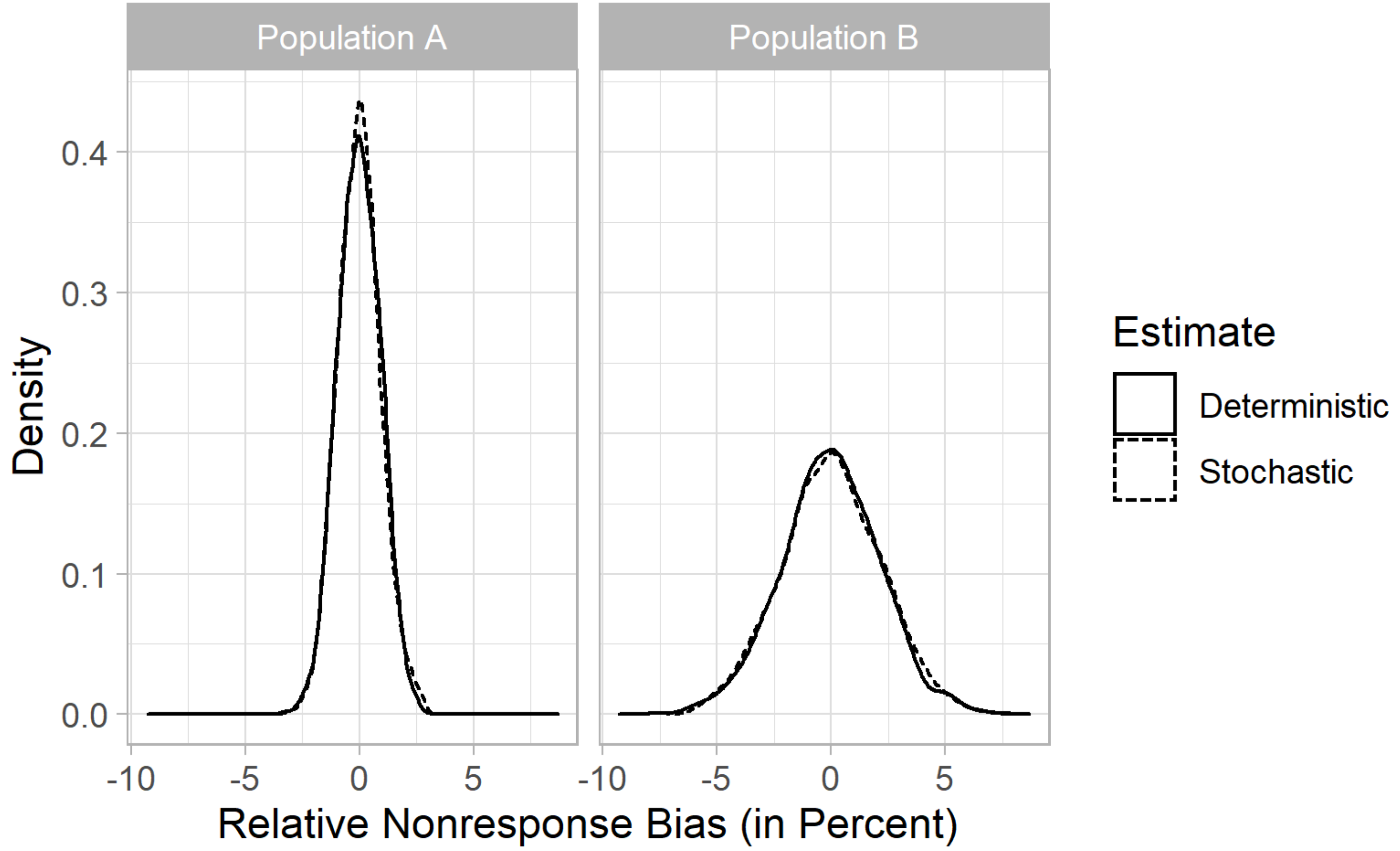
1. How variable are NR bias estimates?
   – RR
   – Clustering

2. How to estimate $Var(bias(\bar{y}_r))$?

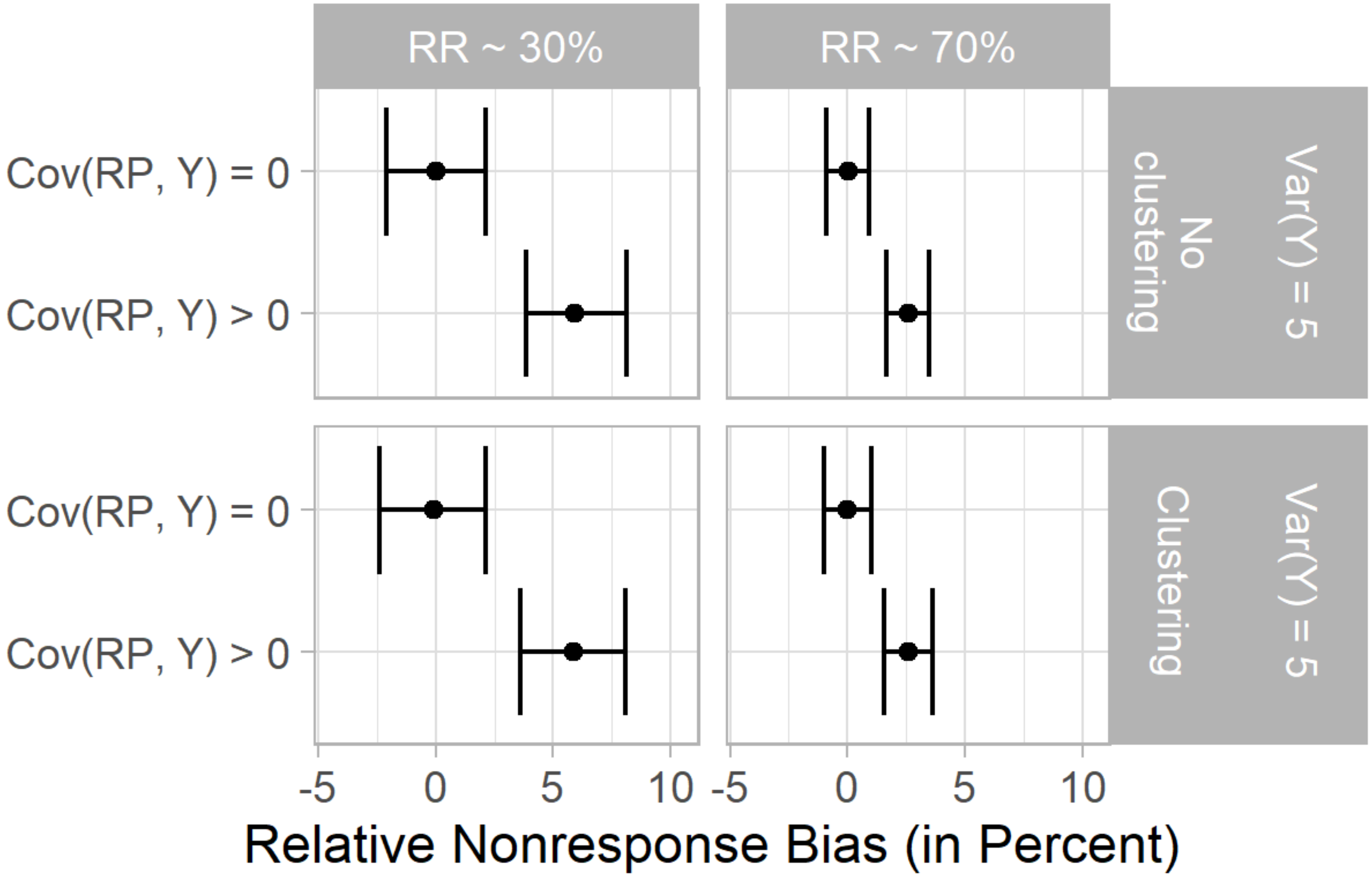
Example Results: Estimated Bias in Simulated Samples

RR ~ 70%
Bias = 0
Unclustered

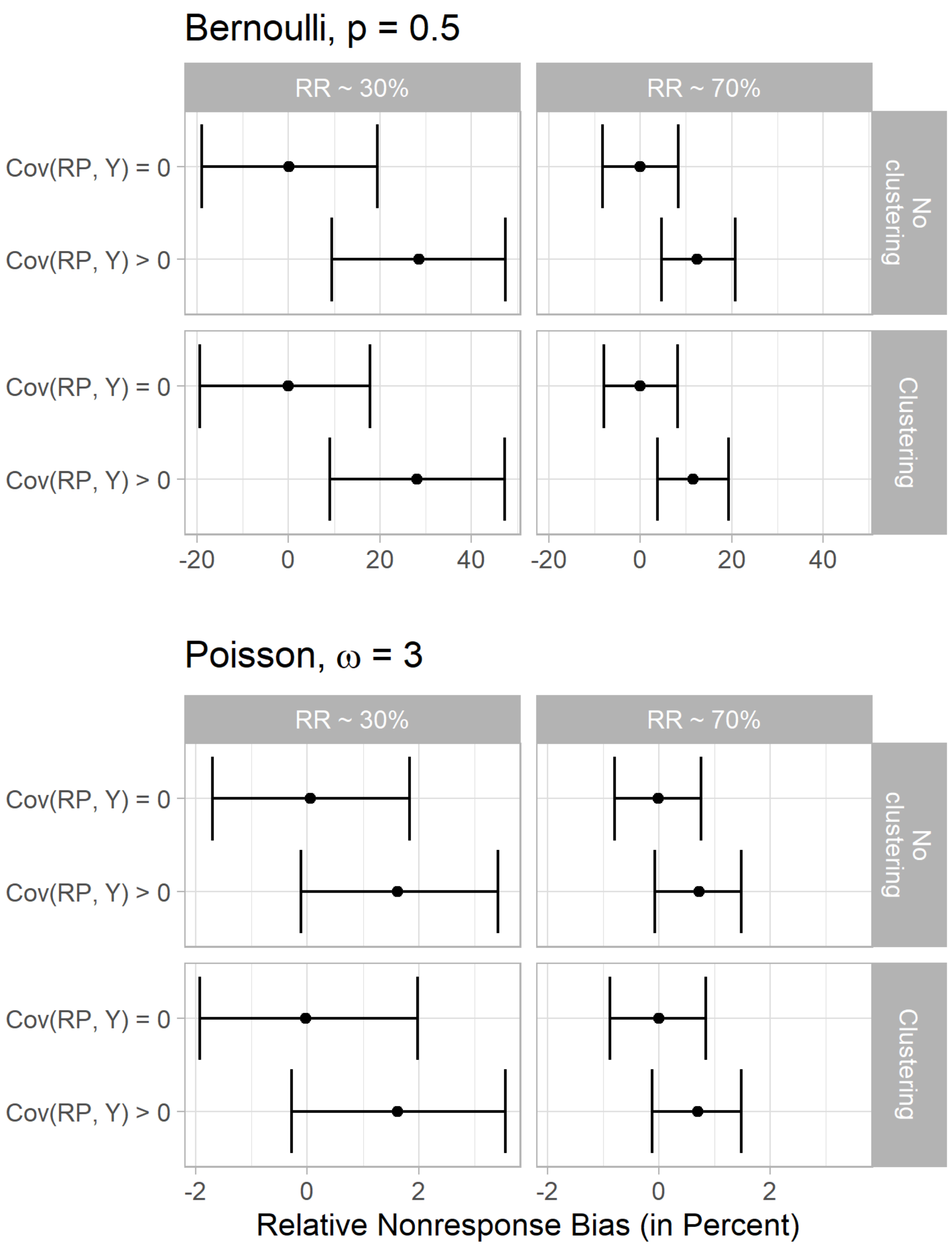
RR ~ 30%
Bias = 0
Clustered



All Populations: Estimated Bias in Simulated Samples



Graphs show 2.5% - 97.5% range



Bernoulli, p = 0.5
RR ~ 30%
RR ~ 70%
No clustering
Clustering
Cov(RP, Y) = 0
Cov(RP, Y) > 0
-20
0
20
40
Poisson, ω = 3
-2
2
Relative Nonresponse Bias (in Percent)

Research Questions

1. How variable are NR bias estimates?
   – RR
   – Clustering

2. How to estimate $Var(bias(\bar{y}_r))$?

Variance Estimation Methods

- Lee (2006) method

$$bias(\bar{y}_r) = (1 - rr) \times (\bar{y}_r - \bar{y}_{nr})$$

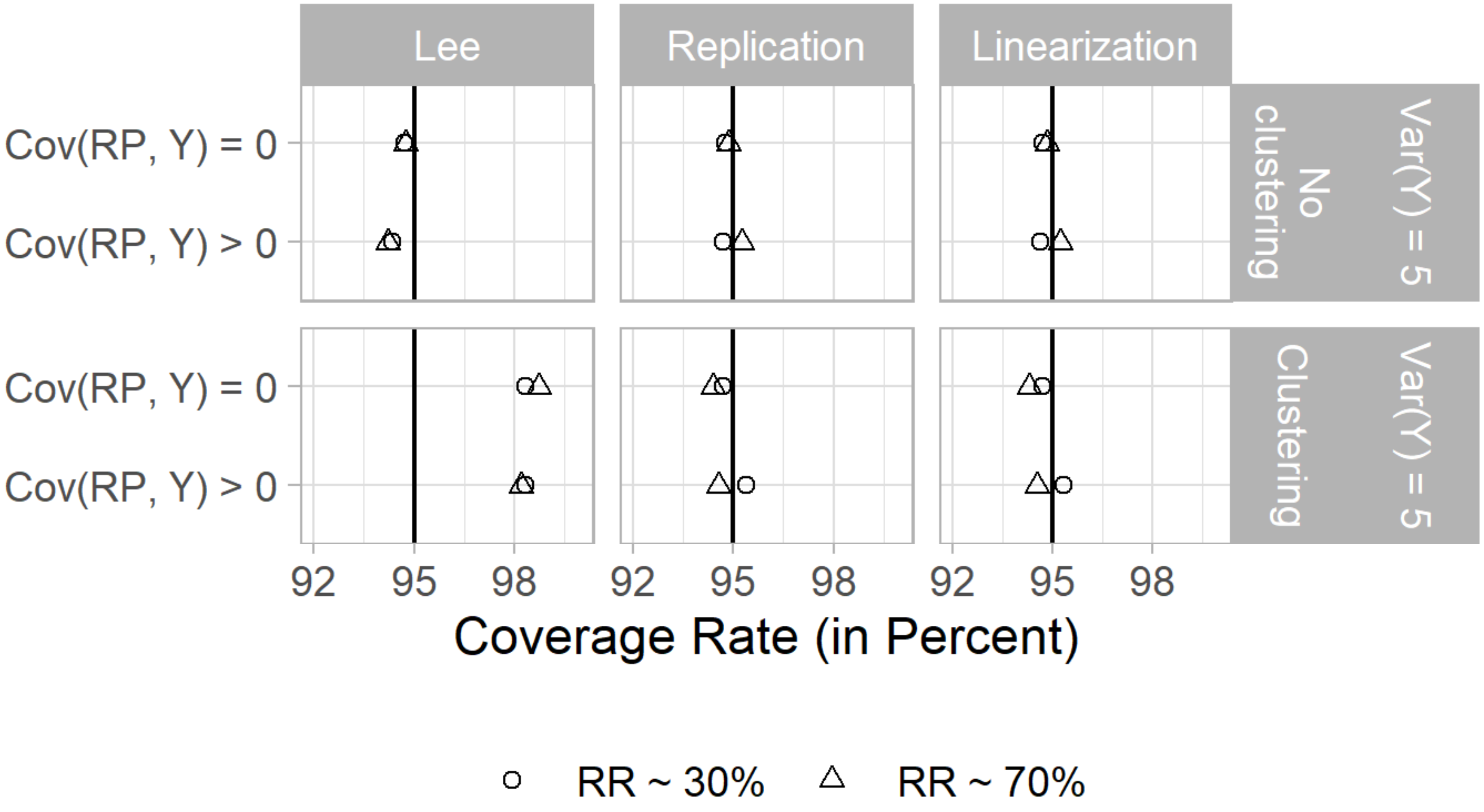
$$Var[bias(\bar{y}_r)] = (1 - rr)^2 \times [Var(\bar{y}_r) + Var(\bar{y}_{nr})]$$

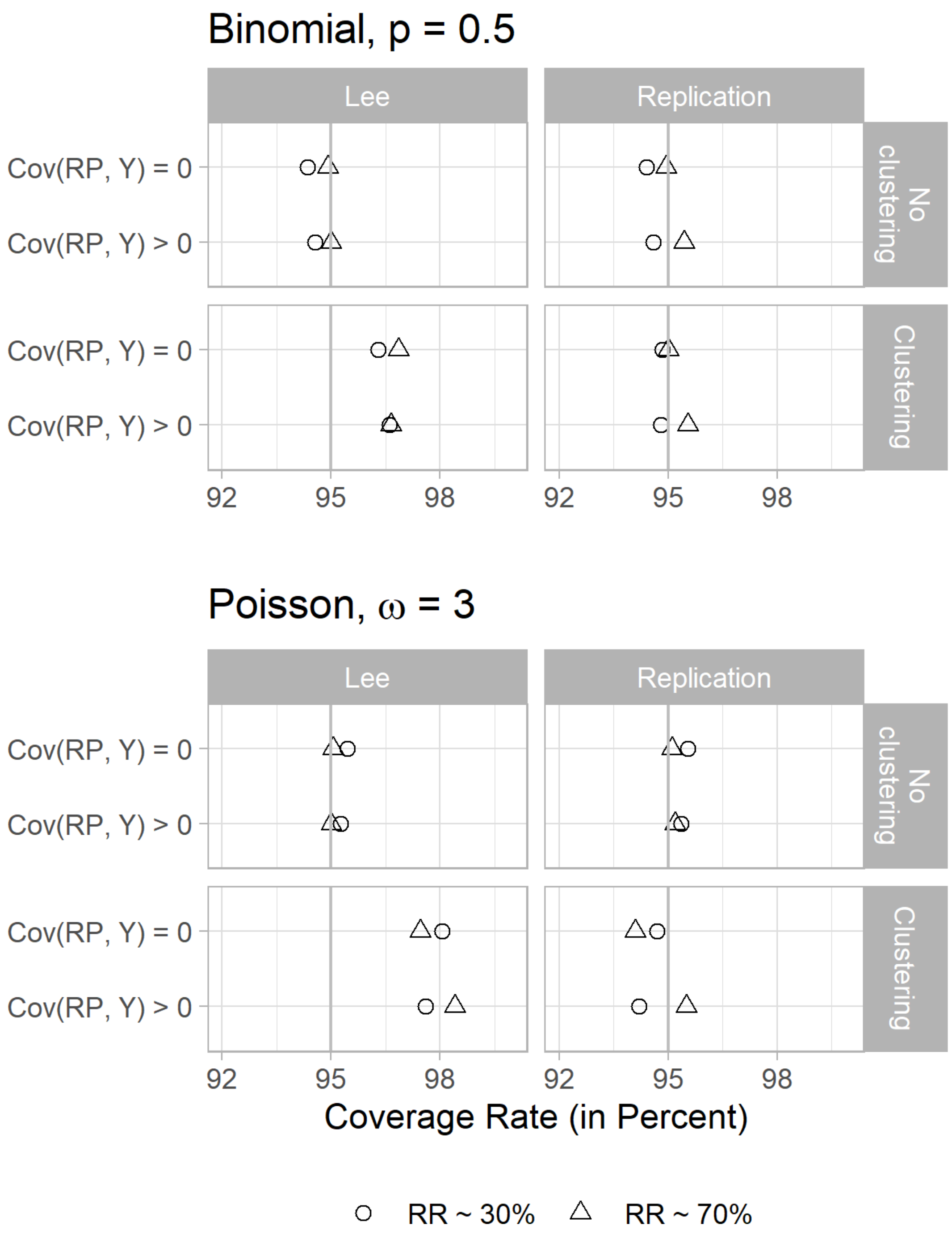
Variance Estimation Methods

- Replication & Linearization

ID	Response?	Base weight	Adjusted weight	Y	Set	Analysis weight
1	Yes	100	140	10.3	1	140
2	Yes	150	190	9.8	1	190
3	No	80	0	8.7	1	0
1	Yes	100	140	10.3	2	100
2	Yes	150	190	9.8	2	150
3	No	80	0	8.7	2	80

Simulation Study – Performance of Three Variance Estimation Methods




Binomial, p = 0.5
Lee
Replication
No clustering
Clustering
Cov(RP, Y) = 0
Cov(RP, Y) > 0
92
95
98
Poisson, ω = 3
Coverage Rate (in Percent)
RR ~ 30%
RR ~ 70%

Application: LISS and GSS

Variable	Type	n	bias	Lee	Repl	Ratio
Hours of TV per weekday	Continuous	623	0.078	0.029	0.030	0.972
Ever driven drunk	Indicator	623	0.013	0.007	0.007	0.986
Immigrants improve NL	Indicator	622	0.011	0.007	0.007	0.987
Approval of Obama	Continuous	622	0.356	0.388	0.389	0.998
Times exercise per week	Count	623	0.003	0.008	0.008	1.004
Times restaurant last year	Count	623	-0.05	0.198	0.197	1.004
Occupational prestige	Continuous	1889	-0.012	0.187	0.204	0.918
Voted in 2004	Indicator	1754	0.023	0.006	0.003	1.725
Hours of TV per day	Continuous	1426	0	0.064	0.031	2.029
Finances better	Indicator	2033	-0.004	0.008	0.003	2.611
Believe in afterlife	Indicator	1803	0.017	0.009	0.003	2.895
Own gun	Indicator	1233	0	0.005	0.001	6.495

Ratio is Lee standard error / replication standard error

Take Aways

- Estimates of NR bias in means vary across samples
- Lee method works well with unclustered samples
- Replication & linearization work with all samples
- Careful when generalizing from 1 sample to another

Thank You

Stephanie Eckman
Fellow, RTI International

Code, paper, slides:
https://osf.io/rbzyd/

seckman@rti.org
http://stepheckman.com
@stephnie

Commands

	Stata (linearization)	R (JK replication)
Describe unclustered data	svyset ID [pweight = v]	dsg <- assvrepdesign(svydesign(ids = ~ID, data = dset, weights = ~v)
Describe clustered data	svyset cluster [pweight = v]	dsg <- assvrepdesign(svydesign(ids = ~cluster, data = dset, weights = ~v)
Estimate bias	svy: mean y, over(set)	bias <- svyby(~y, ~set, dsg, svymean, covmat = TRUE)
Estimate standard error of bias estimate	lincom _b[c.y@1bn.set] – _b[c.y@2.set]	SE(svycontrast(bias, quote(‘1’ – ‘2’)))