Online Supplement: Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys [*]

Carolina Franco[†] Roderick J. A. Little [‡] Thomas A. Louis [§] Eric V. Slud [¶]

August 2, 2018

[*] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[†] US Census Bureau

[‡] University of Michigan

[§] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[¶] US Census Bureau and University of Maryland, College Park

About the Supplement

This online supplement accompanies the article

Franco, C. Little, R. J. A., Louis, T. A., and Slud, E. V. (2018), “Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys, ” *Journal of Survey Statistics and Methodology*

which is referred to throughout this supplement as “the paper.”

Formula number references given in this supplement pertain to the formulas in the paper.

The supplement includes two parts. Section 1 describes the `R` functions and dataset provided for readers in the accompanying workspace `RSupp.RData`. Section 2 includes additional simulation results not included in the paper.

1 R Code for Design Effect and CI Calculations

We describe the use of two `R` functions adapted from those used in the simulations of Section 4 of the paper, `VarKish` and `CIarrFcn`. These functions along with parameter values and data objects used in the illustrations are contained in the supplementary `R` workspace `RSupp.RData`, where function listings and the data objects used below can be found. After explaining the inputs and outputs of the functions in successive subsections, we present a detailed example of the use of these functions similar to the way they were applied in the simulations.

1.1 Function for Survey Variances in Stratified SRS Samples

The `R` function `VarKish` calculates the design-based and ‘Kish-type’ variances for a survey-weighted total of a binary attribute Y_{hki} in the setting of a stratified single-stage cluster-sample (in which all units are taken from each sampled cluster). Some notation is needed, following that of Section 3 and Appendix B of the paper, to explain the data structure of the inputs and the formula numbers corresponding to the outputs. We denote strata by indices $h = 1, \dots, H$, clusters within strata by indices $k = 1, \dots, K_h$, and cluster sizes by M_{kh}. The input data-frame consisting of three columns corresponding to sampled clusters is `Yfr`. The column entries of `Yfr` are, in order: the cluster attribute total Y_{hk+} for each sampled cluster, the stratum label h for the clusters, and the cluster size M_{kh}. The numbers of sampled clusters in stratum h (denoted n_h^C in the paper) are encoded in an input vector `nCvec` with entries `nC[h]`. In this function, the vectors `Kvec` of total numbers K_h of clusters in the strata of the population are assumed known and are also inputs. The primary outputs of the function `VarKish` for these single-stage cluster-sampled data are `Vdsgn`, which is the design-based estimator given by formula (13) of the paper for the variance of $\hat{Y}$, and `VestM`, the superpopulation-model-based variance estimate given by formula (26). Note that as indicated in Section 4 of the paper, in the simulations the estimator that was used for $\hat{\sigma}_h^2$ was $(n_h^C/(n_h^C - 1))\hat{\tau}_h(1 - \hat{\tau}_h)$.

1.2 Function for Confidence Interval Calculations

The `R` function `CIarrFcn` encodes the calculation of all 8 types of confidence intervals studied in the paper. The inputs are two vectors `kstar` and `m` of the same length r, respectively a set of (survey-weighted and scaled) attribute totals and the corresponding effective sample sizes to which they should be referred. In the notations of Sections 2.1 to 2.7 of the paper, the entries of `kstar` correspond to X and entries of `m` correspond to n. In the complex-survey notations of the later sections of the paper, entries of `kstar` correspond to $\hat{n}_{\text{eff}} \cdot \hat{Y}/N$, and entries of m to $\hat{n}_{\text{eff}}$. Neither of the vectors `kstar, m` is required to have integer entries.

The output of the function `CIarrFcn` is an array of dimensions $r \times 5 \times 8$. The first index is the same as the index of vectors `kstar`, `m`, and the third index ranges over the 8 intervals studied in the paper. The second index provides the two-sided confidence interval (CI) endpoints (at the first two index levels “Lowr” and “Upr”), the width, a “Flag” logical indicator that the CI extends outside [0,1] or (in Bayesian intervals) does not contain $\hat{p} = \texttt{kstar}[i]/\texttt{m}[i]$, and a “LoInd” logical indicator (when “Flag” is T) that “Lowr” < 0 or $\hat{p} <$ “Lowr.” Recall that only the Wilson and Uniform intervals are recommended in the paper for general use, with Clopper-Pearson also useful for analysts preferring conservative intervals or Jeffreys for analysts willing to tolerate occasional anti-conservative ones. So the output arrays may be restricted to third index levels 2 through 5.

1.3 Usage Example for the R Functions

We begin by describing the data objects in the `R` work-space `RSupp.RData`. First, `PAR.in` is a parameter list containing arguments

```
> names(PAR.in)
[1] "N"     "J"     "NJvec" "Kvec"  "thvec" "nCvec"
> rbind(NJvec=PAR.in$NJvec, Kvec=PAR.in$Kvec,
            thvec=PAR.in$thvec, nCvec=PAR.in$nCvec)
           [,1]      [,2]      [,3]      [,4]
NJvec 2500.0000 2500.0000 2500.0000 2500.0000
Kvec   500.0000  500.0000  500.0000  500.0000
thvec    0.1875    0.2625    0.3375    0.4125
nCvec    7.0000    9.0000    9.0000    7.0000
```

and `Yfr0` is a 2000×3 data-frame, with the 2000 rows corresponding to the clusters, all of size 5, in a finite population of $N = 10,000$ units divided into 4 strata of 500 clusters each. The first column `Ycl` consists of the 2000 cluster-totals of binary attribute values Y; the second column `jvec` contains the stratum labels 1:4 of the clusters, and the third column `csize` consists of the cluster sizes, in this case all 5's.

```
> table(Yfr0$Ycl)
  0   1   2   3   4   5
705 466 345 242 154  88
```

In this instance, the finite-population total of the Y's is 2938, with population (unit) average equal to the proportion 0.2938.

One further object in `RSupp.RData` is `indx0`, a 32×5 array, each column of which is the set of indices $1 : 2000$ of a stratified sample of 32 clusters, in which `PAR.in$nCvec[j]` clusters are drawn from stratum `j`. This array provides indices for 5 independent samples, although only the first sample is used below. Then `Yfr0[indx0[,1],]` is a data-frame of cluster-totals, stratum labels and cluster sizes for the first stratified random sample from the finite population represented in `Yfr0`. In practice as opposed to simulations, the user would have access only to the sample.

```
> table(Yfr0$jvec[indx0[,1]])
1 2 3 4
7 9 9 7
```

We examine the outputs from applying the function `VarKish` on this sample dataset:

```
> tmp0 = VarKish(Yfr0[indx0[,1],], 4, PAR.in$nCvec, PAR.in$Kvec)
> unlist(tmp0[c(1:3,6)])
       Vdsgn        VestM         Yest       rhohat
1.562614e+05 1.479138e+05 2.246032e+03 9.423208e-02
> rbind(tauhat=tmp0$tauhat, sighsq=tmp0$sighsq)
               1         2         3         4
```

```
tauhat 0.1714286 0.2888889 0.2666667 0.1714286
sighsq 0.1462185 0.2101010 0.2000000 0.1462185
```

Here, the estimated total was `Yest`$= 2246.03$, where recall that the actual finite-population total was 2938 and proportion was 0.2930. The design standard error was $\sqrt{156261.4} = 395.3$, and the estimated standard error according to the ‘Kish method’ (variance formula (26)) was 384.6. The estimated ICC was $\hat{\rho} = 0.0942$. The estimated stratum proportions $\hat{\tau}_h$ and variances $\hat{\sigma}^2_h$ calculated by VarKish are also given above.

Using the same data and `VarKish` output, we define two estimated sample sizes (`mstar`) corresponding to the `Vdsgn` and `VestM` and corresponding sets of scaled counts `kstar`:

```
> Yest = tmp0$Yest; n=32*5; N=PAR.in$N; f=n/N
> mstar = Yest*(N-Yest)*(1-f)/c(tmp0$Vdsgn, tmp0$VestM)
> kstar = mstar*Yest/N
> round(rbind(mstar=mstar, kstar=kstar), 2)
        [,1]   [,2]
mstar 109.67 115.86
kstar  24.63  26.02
```

We continue by calculating the full set of confidence intervals for each of the pairs `(kstar[1], mstar[1])` associated with estimates effective sample sizes using formula (12), and `(kstar[2], mstar[2])` with estimated effective sample sizes using formula (18):

```
> CIarrFcn(kstar,mstar,0.05)
, , Wald
       Lowr       Upr     Width Flag LoInd
1 0.1464986 0.3027077 0.1562091    0     0
2 0.1486134 0.3005929 0.1519795    0     0
```

```
, , JeffPr

       Lowr       Upr     Width Flag LoInd
1 0.1543083 0.3092113 0.1549030    0     0
2 0.1559952 0.3067710 0.1507758    0     0

, , UnifPr

       Lowr       Upr     Width Flag LoInd
1 0.1567177 0.3115236 0.1548060    0     0
2 0.1582818 0.3089685 0.1506867    0     0

, , ClPe

       Lowr       Upr     Width Flag LoInd
1 0.1504228 0.3141947 0.1637719    0     0
2 0.1522975 0.3114811 0.1591837    0     0

, , Wils

       Lowr       Upr     Width Flag LoInd
1 0.1565880 0.3112585 0.1546705    0     0
2 0.1581603 0.3087224 0.1505621    0     0

, , AgCo

       Lowr       Upr     Width Flag LoInd
1 0.1560473 0.3117992 0.1557519    0     0
2 0.1576598 0.3092229 0.1515631    0     0

, , Assqr
```

```
       Lowr       Upr     Width Flag LoInd
1 0.1539044 0.3098104 0.1559059    0     0
2 0.1556235 0.3073243 0.1517008    0     0

, , Logit
       Lowr       Upr     Width Flag LoInd
1 0.1561032 0.3120470 0.1559438    0     0
2 0.1577096 0.3094462 0.1517366    0     0
```

In this `R` output array, the headers “Wald”, . . ., “Logit” index the type of confidence interval, the row-indices 1 and 2 refer to the two types of design-effect computations provided in the supplied `mstar, kstar` input arguments (respectively, 1 for `Vdsgn` the standard design-effect calculation based on SRS single-stage cluster sampling, and 2 for `VestM` the design-effect calculation based on the ‘Kish method’ of the paper). Each interval was calculated as in Section 5 with `Lowr, Upr` the interval endpoints and `Width` the difference between them. Two other columns `Flag, LoInd` were all zeros in this example. Their meaning is as follows: for all intervals other than “JeffPr” and “UnifPr”, `Flag` is the indicator of the event that the interval is not a subset of $[0, 1]$, and `LoInd` is the indicator that the left-hand endpoint of the interval is negative. For the intervals “JeffPr” and ”UnifPr”, `Flag` is the indicator that the point estimator [which is $(k^* + 0.5)/(m^* + 1)$ for “JeffPr” and $(k^* + 1)/(m^* + 2)$ for “UnifPr”] falls outside the interval while `LoInd` indicates that the point estimator is to the left of the lower interval endpoint. Here the scaled counts `kstar`$= k^*$ and effective sample sizes `mstar`$= m^*$ respectively play the roles of the counts X and sample sizes n in the confidence interval formulas of Section 2 of the paper.

2 Additional Simulation Results

All exhibits are based on the simulation results for all intervals excluding the Wald, as the Wald was shown to perform poorly in the paper. Details of the simulation are provided there. Tables 1-3 and Figures 1-6 also exclude the Logit interval, which is treated separately in Figures 7-11.

Note the abbreviations: `JeffPr` refers to Jeffreys interval, `UniPr` to Uniform, `ClPe` to Clopper-Pearson, `Wils` to Wilson, `AgCo` to Agresti-Coull, and `Assqr` to Arcsine Square Root interval.

Tables 1-3 show the percentage of instances over the simulation configurations in which the intervals have coverage below the thresholds of 93.5%, 94%, and 94.5%, respectively, by n, c, and ICC.

Figures 1-2 contrasts the ratios of non-coverages with the ratios of widths for the Kish to the DP methods, for $c = 3$ and $c = 5$ respectively. The ratios are multiplied by 100 for ease of exposition. The case $c = 1$ is discussed in the paper and the case $c = 7$ is displayed in the paper (Figure 3).

Figures 3-4 plot the coverage using the Kish method vs. the DP method for each of the intervals and for each point in our simulation configuration, for $c = 3$ and $c = 5$ respectively. The paper discussed $c = 1$ and displays $c = 7$ (Figure 4).

Figures 5-6 plot the ratio of each interval width to that of Clopper-Pearson (since that CI is typically the widest and has highest coverage) versus the non-coverage, plotting a separate panel for each level of clustering $c = 3$ and 7, and within each c, for each overall proportion θ. The cases $c = 1$ and $c = 5$ are shown in the paper (Figures 5 and 6).

Figures 7-11 were added for completeness to illustrate the performance of the Logit interval. Figure 7 shows the coverage of each of the CI's using the design-based estimate of the effective sample size (left panels) and the true effective sample size (right panels). Figures 8-11 are analogous to Figures 5-6, but they add the Logit interval to the mix, plotted in cyan, and contrast its performance to the Agresti-Coull interval, in blue, the Clopper-Pearson, in black, and the others, in gray. They show the similar performance of the Logit interval to the Agresti-Coull, with some cases of very high interval width.

c	1				3			5			7		
n	ICC	0.25	0.1	0.001	0.25	0.1	0.001	0.25	0.1	0.001	0.25	0.1	0.001
30	Design	3	4	3	40	42	40	NA			NA		
	Kish	3	4	3	0	0	0						
	DP	0	1	0	0	0	0						
40	Design	0	0	0	44	39	29	NA			NA		
	Kish	0	0	0	0	0	0						
	DP	0	0	0	0	0	0						
50	Design	0	0	1	32	22	19	NA			NA		
	Kish	0	0	1	0	0	0						
	DP	0	0	0	1	0	0						
84	Design	0	0	0	31	24	6	60	61	29	69	71	58
	Kish	0	0	0	3	1	0	1	0	0	6	0	0
	DP	0	0	0	4	1	0	6	1	1	15	0	0
196	Design	0	0	0	4	0	0	31	10	0	68	40	19
	Kish	0	0	0	3	0	0	11	0	0	25	1	0
	DP	0	0	0	1	0	0	12	0	0	32	6	0
280	Design	0	0	0	4	0	0	19	3	0	42	10	0
	Kish	0	0	0	1	0	0	10	0	0	24	0	0
	DP	0	0	0	1	0	0	10	0	0	28	0	0

Table 1: Percentage of times that the coverage is below 93.5% for a 95% nominal coverage over simulation configurations and over all intervals excluding the Wald and Logit

c	1				3			5			7		
n	ICC	0.25	0.1	0.001	0.25	0.1	0.001	0.25	0.1	0.001	0.25	0.1	0.001
30	Design	12	14	12	52	52	46	NA			NA		
	Kish	12	14	12	7	4	4						
	DP	8	8	7	6	4	4						
40	Design	10	11	10	56	51	49	NA			NA		
	Kish	10	11	10	8	5	5						
	DP	8	6	7	10	8	8						
50	Design	11	13	10	50	44	43	NA			NA		
	Kish	11	13	10	12	4	4						
	DP	7	10	5	11	7	6						
84	Design	8	10	12	51	43	37	76	71	62	79	79	69
	Kish	8	10	12	13	12	6	15	10	5	21	10	4
	DP	8	10	10	19	13	10	21	13	10	31	18	7
196	Design	6	6	7	31	10	11	51	29	20	79	67	49
	Kish	6	6	7	13	8	4	25	7	2	45	13	2
	DP	6	6	7	14	8	7	31	11	7	46	19	7
280	Design	5	4	4	14	7	5	40	15	5	62	30	17
	Kish	5	4	4	11	4	1	24	8	0	38	6	0
	DP	4	4	4	13	5	4	21	10	4	40	8	5

Table 2: Percentage of times that the coverage is below 94% for a 95% nominal coverage over simulation configurations and over all intervals excluding the Wald and Logit

c	1				3			5			7		
n	ICC	0.25	0.1	0.001	0.25	0.1	0.001	0.25	0.1	0.001	0.25	0.1	0.001
30	Design	13	20	14	55	57	50	NA			NA		
	Kish	13	20	14	8	4	4						
	DP	10	12	10	8	6	4						
40	Design	13	14	13	57	54	61	NA			NA		
	Kish	13	14	13	10	5	6						
	DP	8	8	10	15	14	8						
50	Design	15	17	17	51	51	48	NA			NA		
	Kish	15	17	17	15	4	5						
	DP	12	11	8	15	10	7						
84	Design	15	19	24	69	70	56	77	81	69	79	82	74
	Kish	15	19	24	29	15	8	26	11	6	30	10	5
	DP	11	12	17	36	31	13	27	30	10	42	26	10
196	Design	14	15	12	43	29	29	69	60	52	89	81	79
	Kish	14	15	12	23	11	5	42	10	4	58	19	2
	DP	8	13	11	27	12	11	45	10	11	64	33	17
280	Design	7	5	10	20	11	12	56	49	11	79	60	40
	Kish	7	5	10	14	8	4	40	8	1	59	14	1
	DP	7	4	7	14	8	5	42	15	5	51	27	7

Table 3: Percentage of times that the coverage is below 94.5% for a 95% nominal coverage over simulation configurations and over all intervals excluding the Wald and Logit

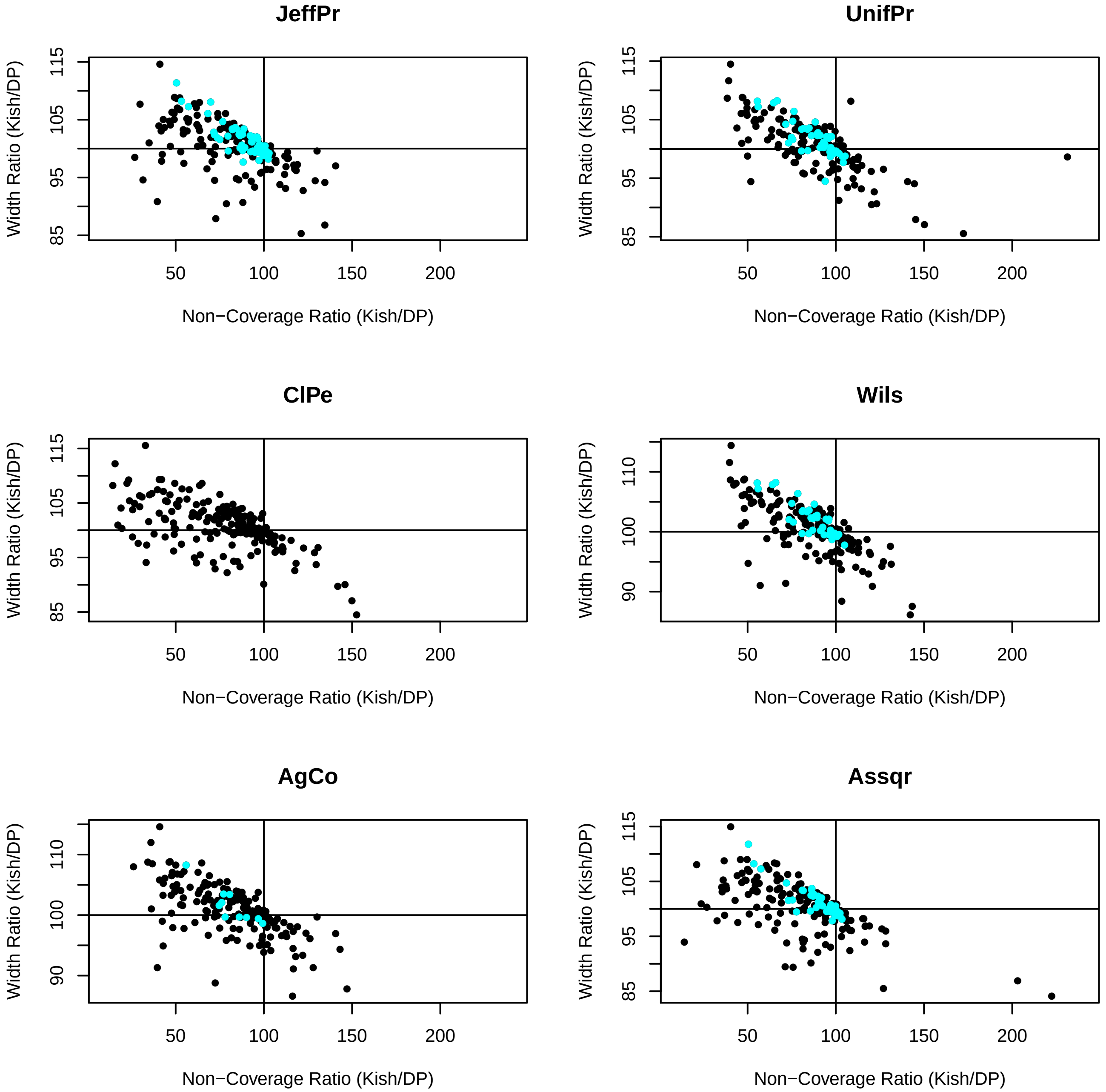

Figure 1: Comparison of metrics for Kish versus DP n_{eff} methods, for each of 6 CIs under 216 simulation settings with $c = 3$. Plotted points are: y = 100 times ratio of widths for Kish n_{eff} method over DP, versus x = 100 times ratio of non-coverage for Kish n_{eff} method over non-coverage for DP. Points with below-nominal DP coverage plotted in cyan. Vertical line indicates coverage ratio 1, horizontal line width-ratio 1

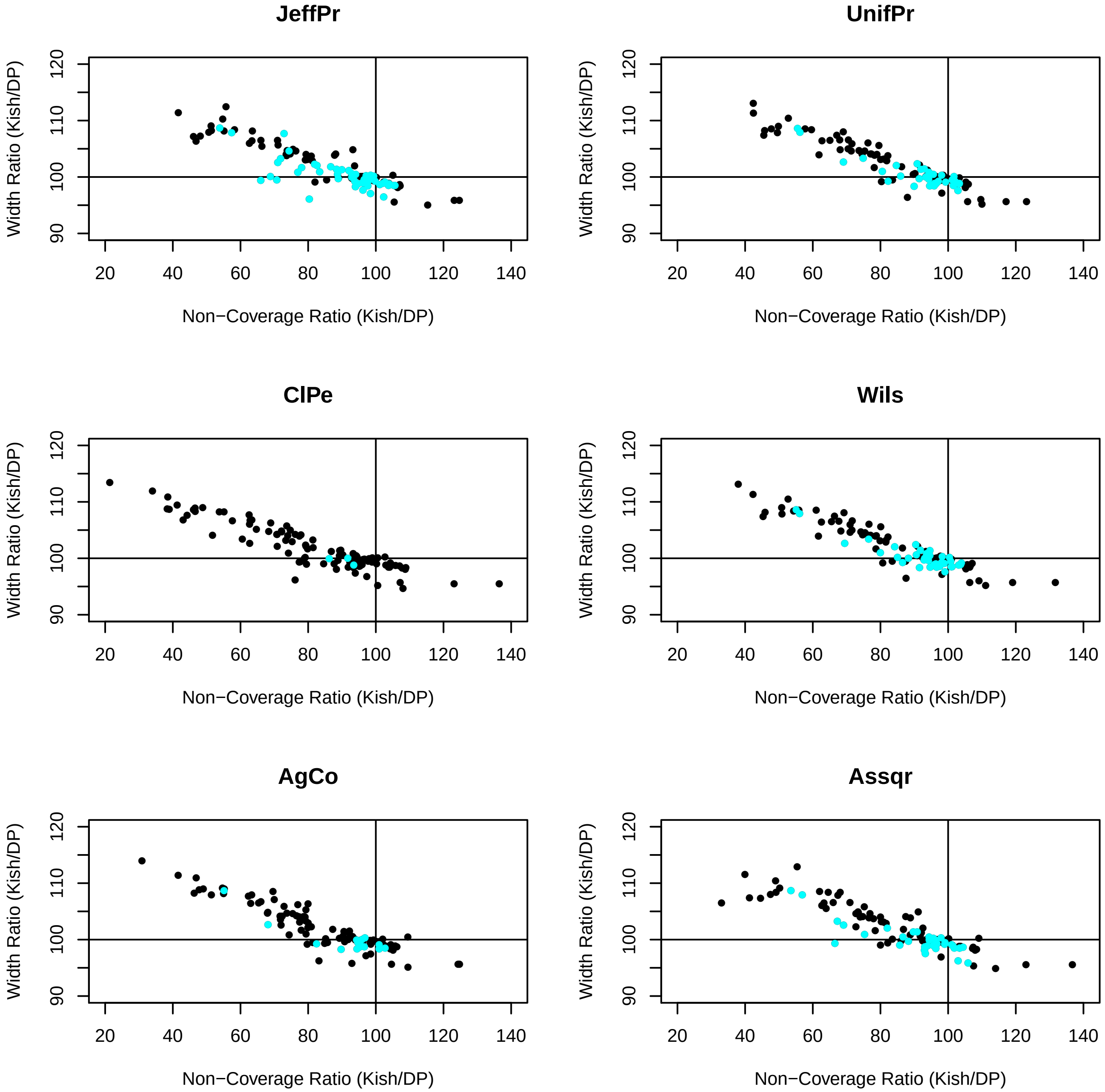

Figure 2: Comparison of metrics for Kish versus DP n_{eff} methods, for each of 6 CIs under 108 simulation settings with $c = 5$. Plotted points are: y = 100 times ratio of widths for Kish n_{eff} method over DP, versus x = 100 times ratio of non-coverage for Kish n_{eff} method over non-coverage for DP. Points with below-nominal DP coverage plotted in cyan. Vertical line indicates coverage ratio 1, horizontal line width-ratio 1

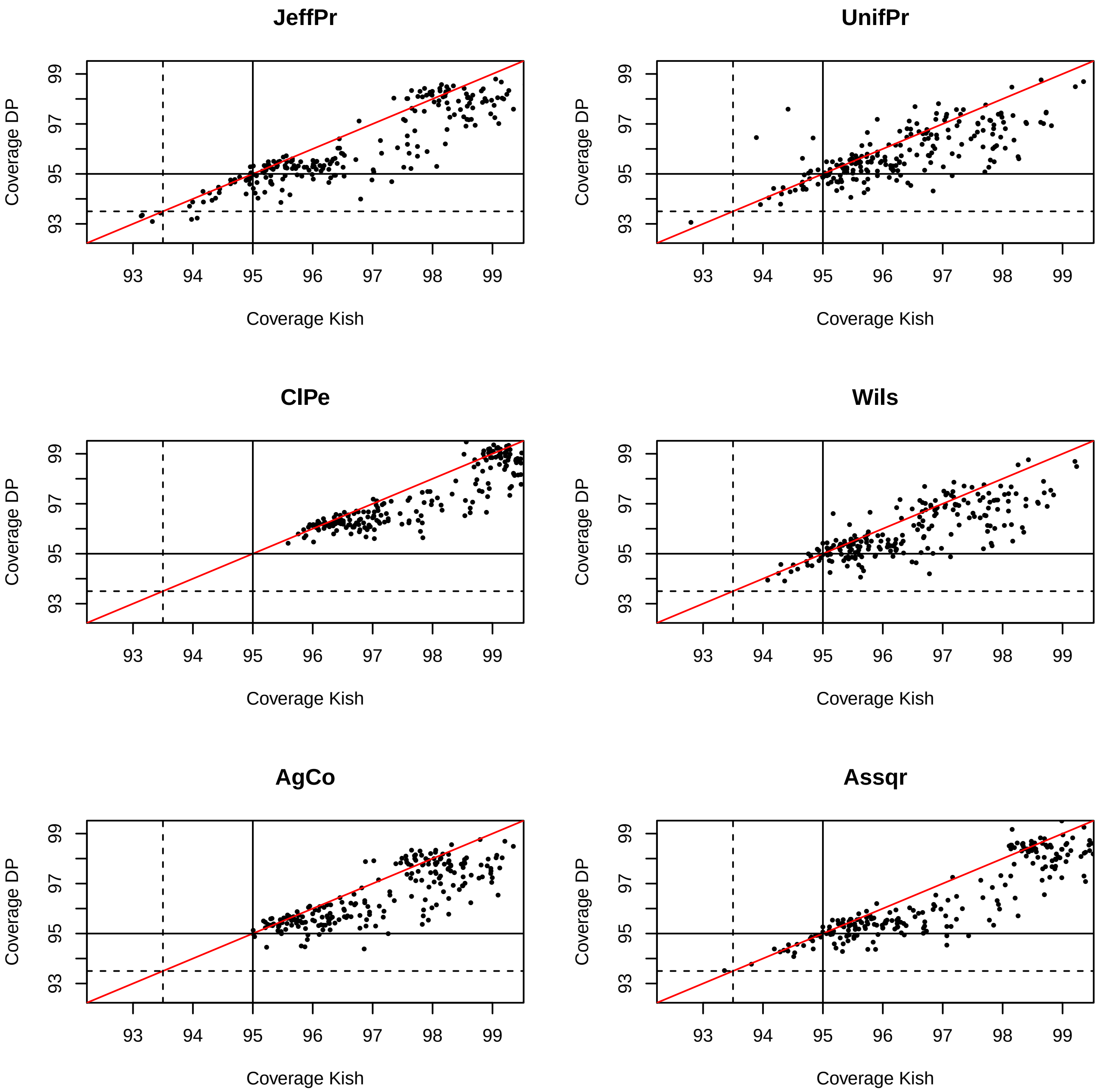

Figure 3: Coverage for 6 CI types, scaled by 100, for the Kish and DP n_{eff} methods in 108 simulation configurations with $c = 3$. Equal coverage is indicated by red $45°$ line, nominal (95%) coverage by black solid lines, and extreme (93.5%) undercoverage by black dashed lines.

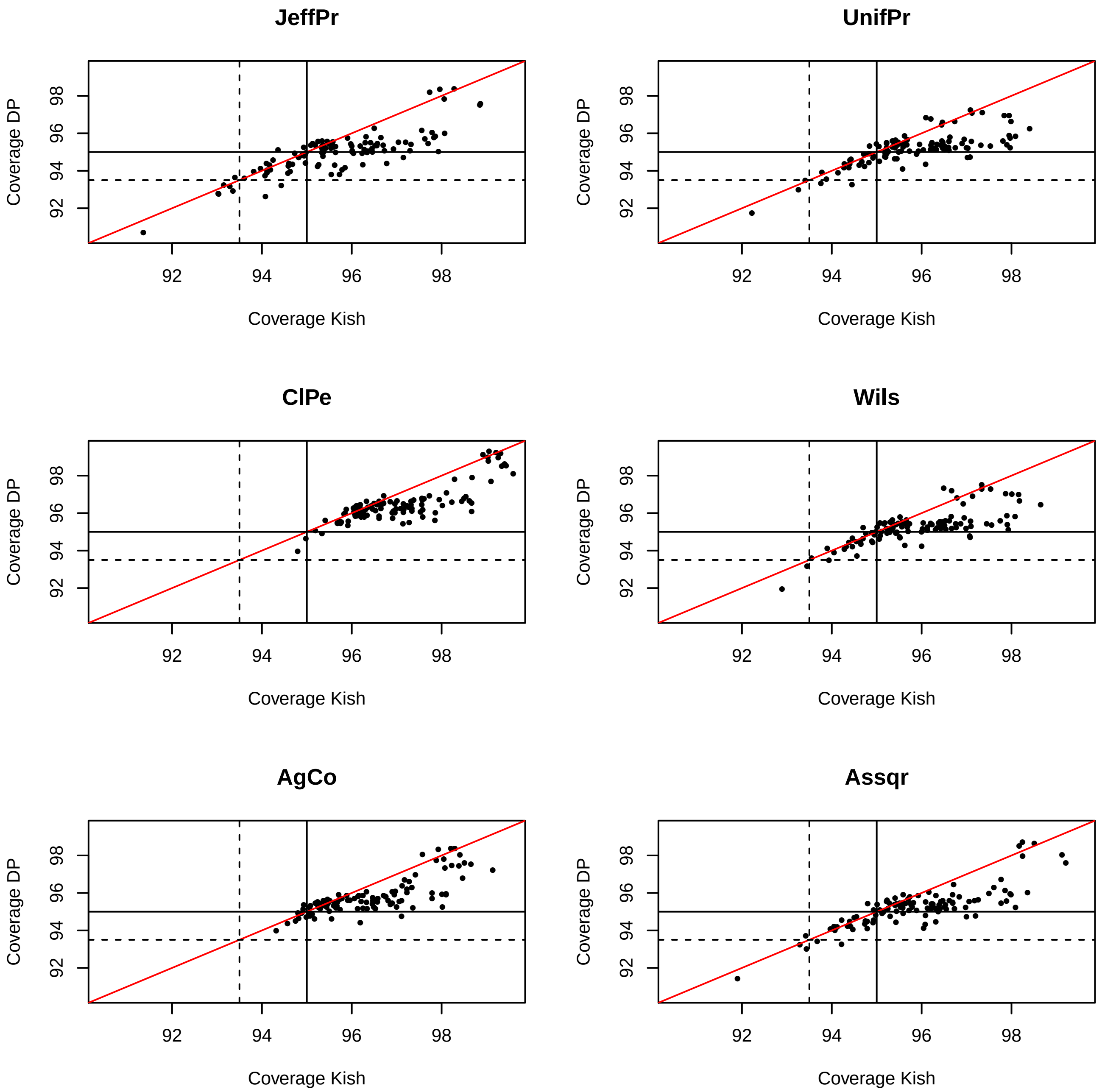

Figure 4: Coverage for 6 CI types, scaled by 100, for the Kish and DP n_{eff} methods in 108 simulation configurations with $c = 5$. Equal coverage is indicated by red $45°$ line, nominal (95%) coverage by black solid lines, and extreme (93.5%) undercoverage by black dashed lines.

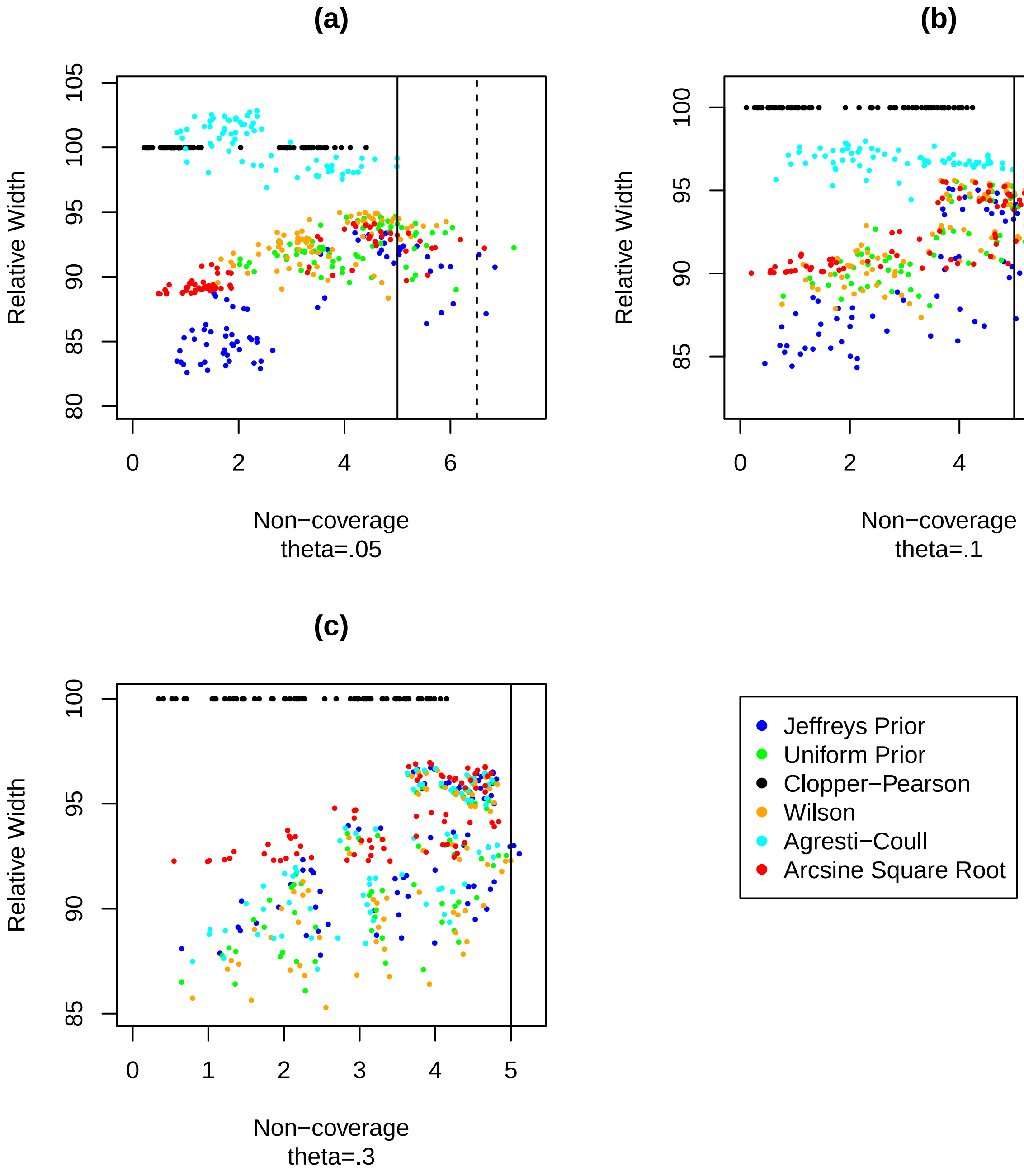

Figure 5: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull, and Arcsine Square Root to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 72 configurations with $c = 3$ and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. Solid vertical line represents nominal non-coverage. Dotted line, given for reference, represents undercoverage of 1.5 percentage points

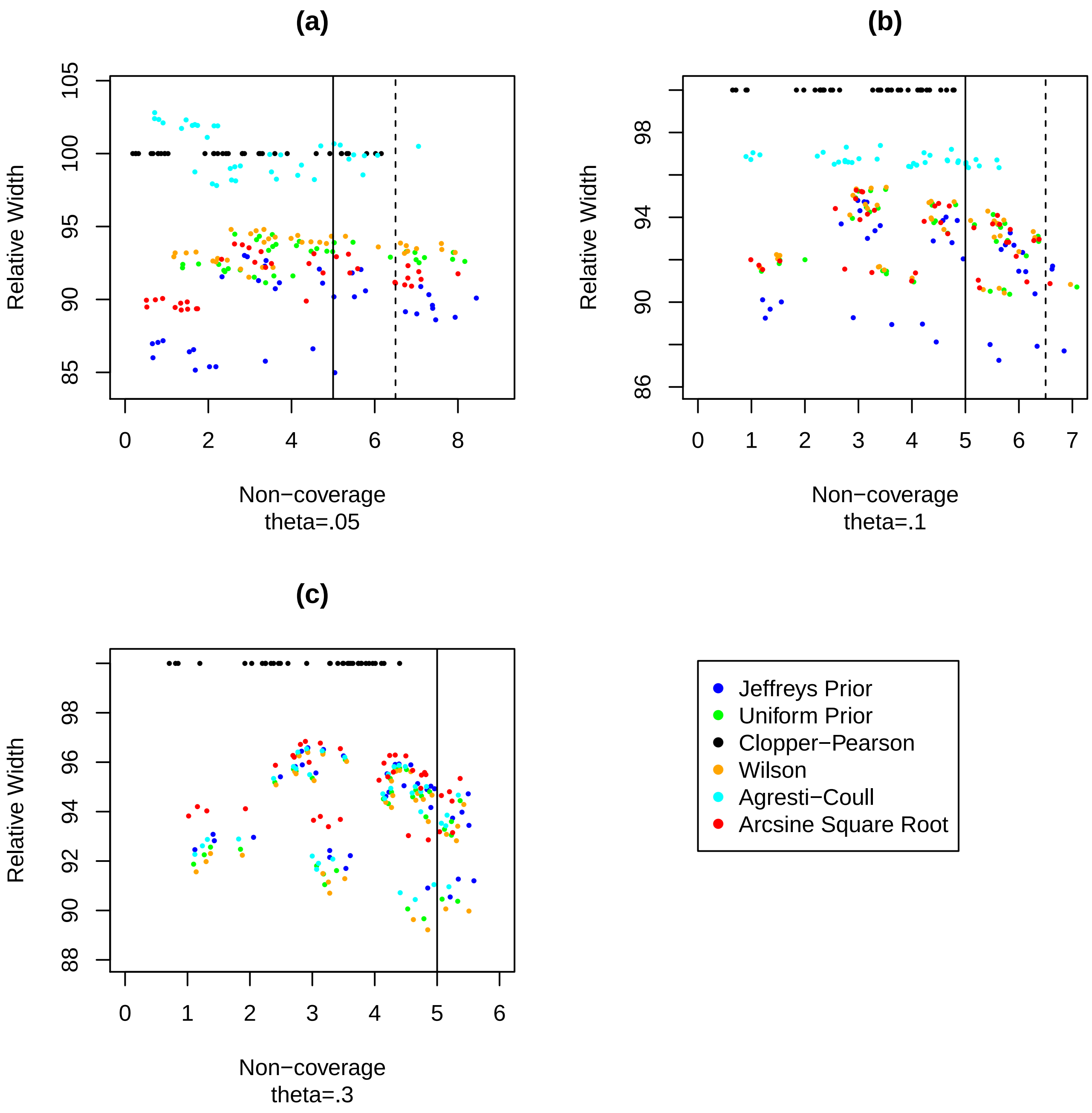

Figure 6: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull, and Arcsine Square Root to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 36 simulation configurations with $c = 7$ and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. Solid vertical line represents nominal non-coverage. Dotted line, given for reference, represents under-coverage of 1.5 percentage points.

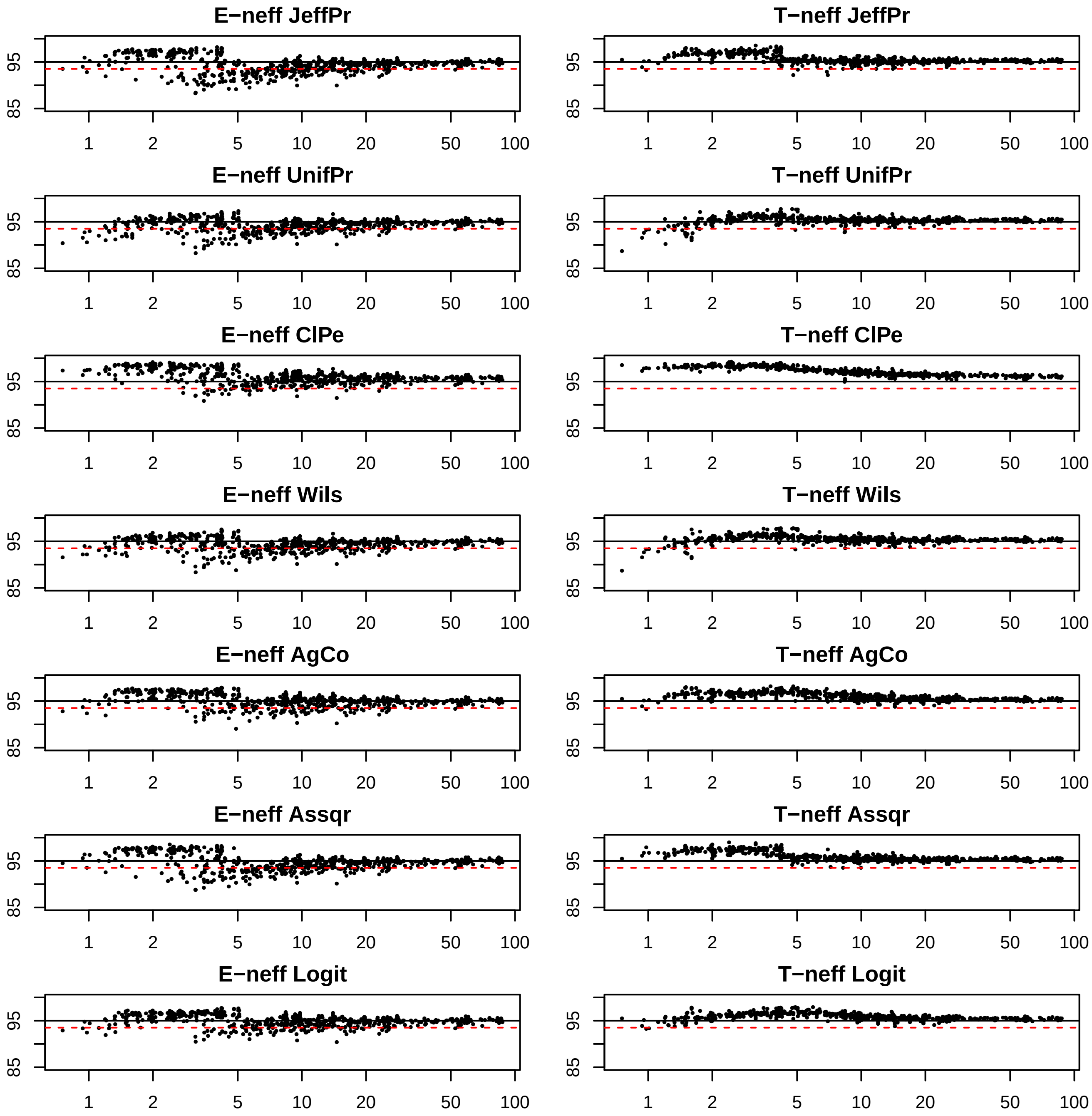

Figure 7: Left Panels: Coverage of 7 CIs using design-based effective sample size estimate, plotted against effective expected number of successes ($n_{\text{eff}} \cdot \theta$, plotted on the log scale), for each simulation configuration. `JeffPr` refers to Jeffreys interval, `UniPr` to Uniform, `ClPe` to Clopper-Pearson, `Wils` to Wilson, `AgCo` to Agresti-Coull, and `Assqr` to Arcsine Square Root interval. Solid line at 95 represents nominal coverage, and dashed line at 93.5 undercoverage by 1.5%. Right panels: Analogous to left panels, using the true effective sample size instead of the design-based estimated effective sample size

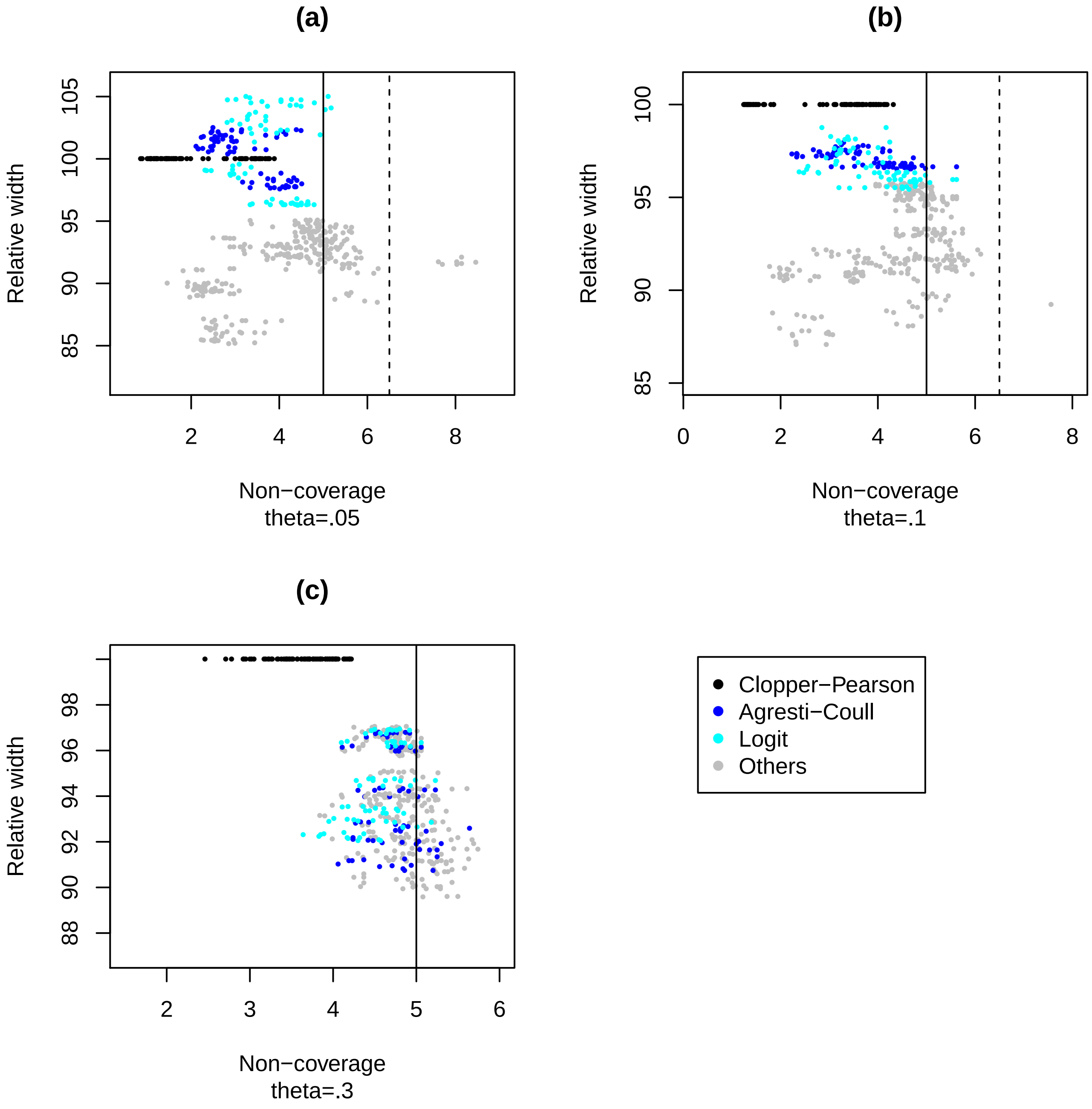

Figure 8: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull, and Arcsine Square Root interval to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 72 configurations with no clustering ($c = 1$) and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. Solid vertical line represents nominal non-coverage. Dotted line, given for reference, represents undercoverage of 1.5 percentage points.

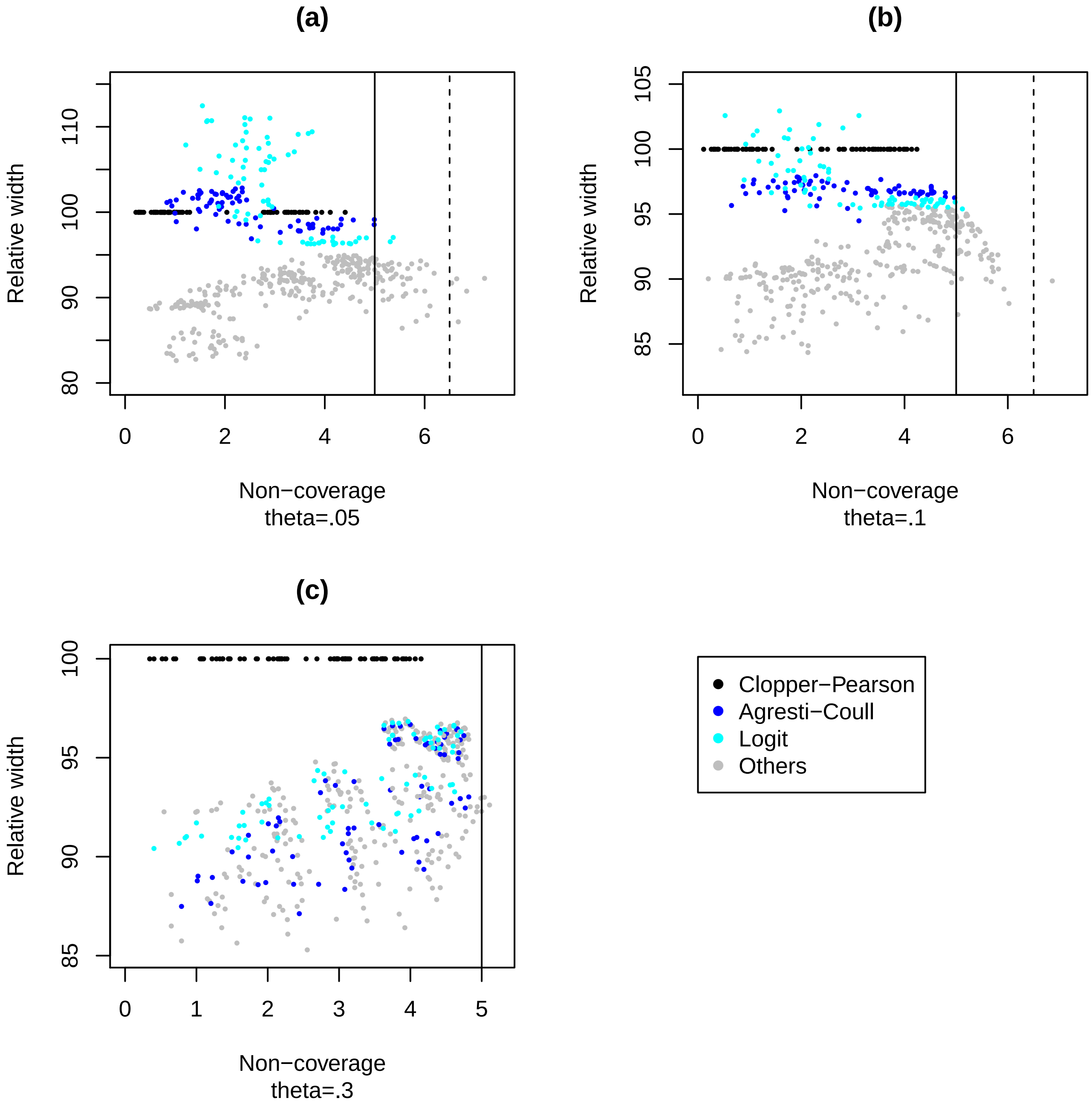

Figure 9: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull, and Arcsine Square Root to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 72 simulation configurations with $c = 3$ and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. Solid vertical line represents nominal non-coverage. Dotted line, given for reference, represents under-coverage of 1.5 percentage points.

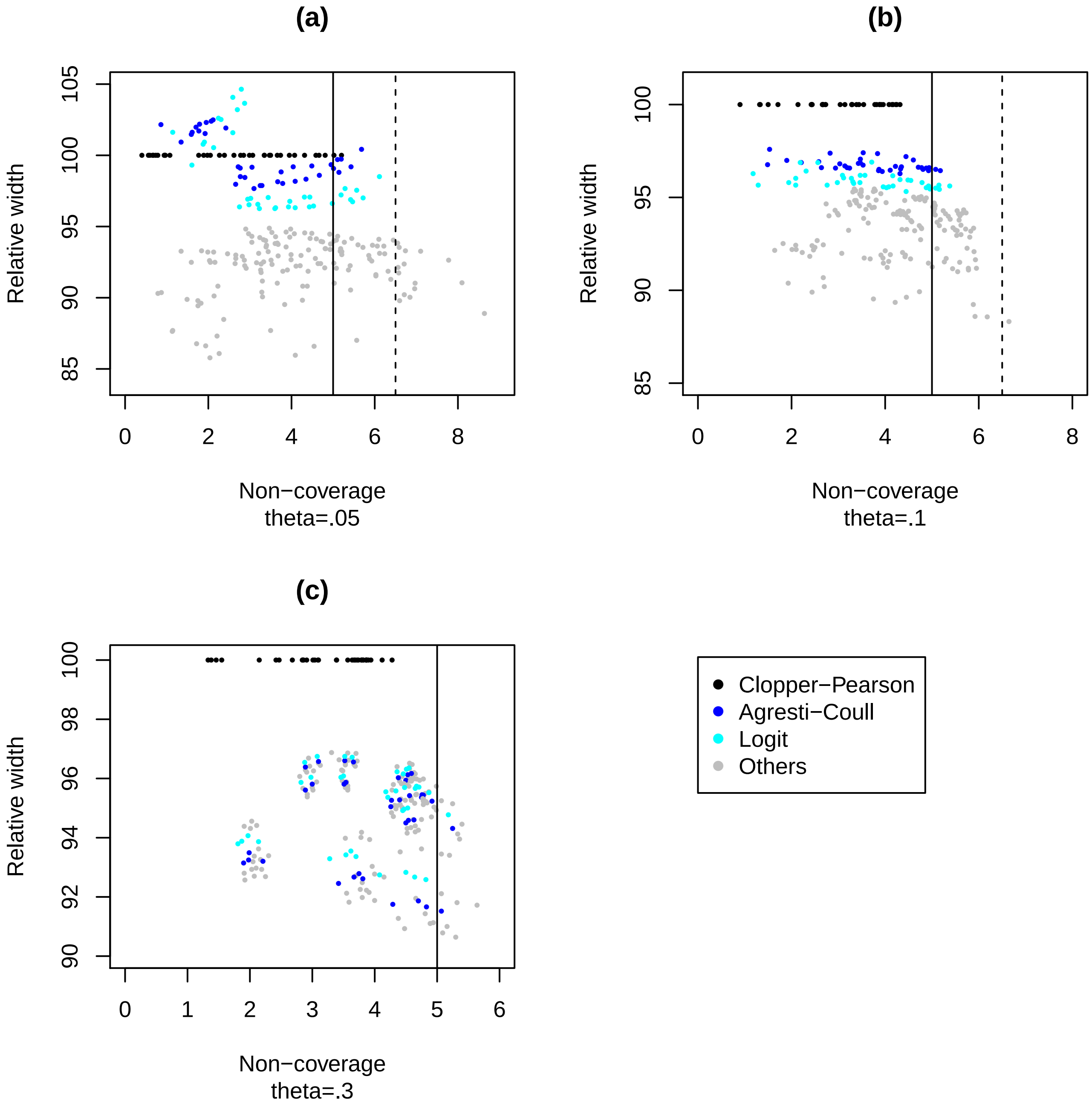

Figure 10: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull, and Arcsine Square Root to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 36 simulation configurations with $c = 5$ and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. Solid vertical line represents nominal non-coverage. Dotted line, given for reference, represents under-coverage of 1.5 percentage points.

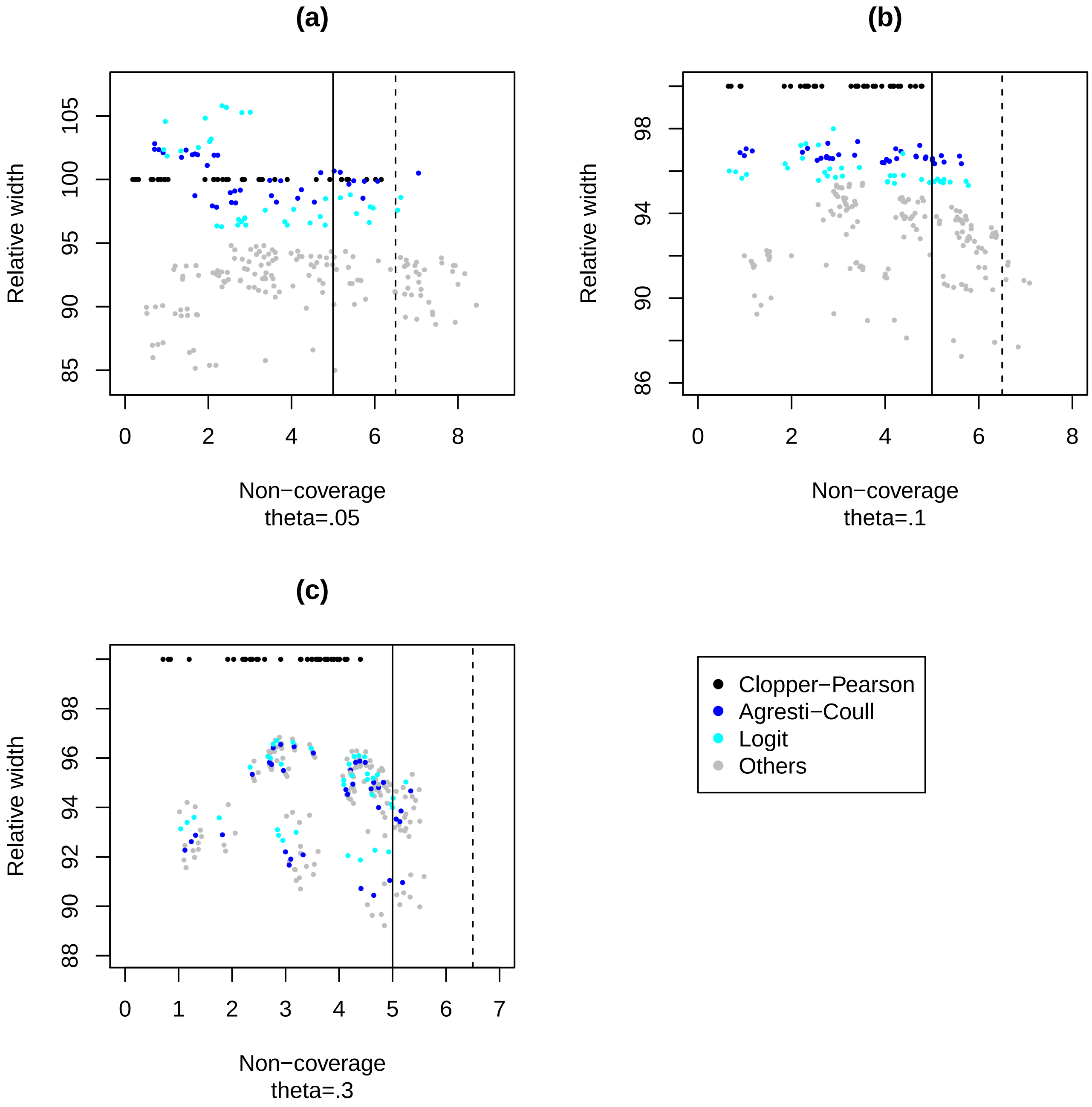

Figure 11: Relative width (ratio of width of Jeffreys, Uniform, Clopper-Pearson, Wilson, Agresti-Coull, and Arcsine Square Root to that of the Clopper-Pearson, multiplied by 100) vs non-coverage for 36 simulation configurations with $c = 7$ and (a) $\theta = 0.05$, (b) $\theta = 0.1$, (c) $\theta = 0.3$. Solid vertical line represents nominal non-coverage. Dotted line, given for reference, represents under-coverage of 1.5 percentage points.