Inference of Domain Parameters Through an Automatic Adjustment of Degrees of Freedom

Sixia Chen and Tom Krenzke

Westat
1600 Research Blvd, Rockville, MD, 20850
SixiaChen@westat.com, TomKrenzke@westat.com

Abstract

In this paper, we explore automatic adjustments of degrees of freedom that can be made with inference for small domain parameters to obtain confidence intervals with coverage probabilities closer to the nominal values. A paired jackknife replication variance estimator is evaluated for the inference of domain parameters under a multistage complex sampling design. The degrees of freedom are adjusted according to the membership of sampling units. The proposed method is compared with traditional approaches for approximating degrees of freedom for jackknife replication variance estimators that do not take the distribution of the domain across the primary sampling units into consideration. The proposed method can be effectively applied to online analytic systems (OAS) that produce results in real-time. A limited simulation study based on 2011 National Health Interview Survey (NHIS) public use data is presented in the paper, which evaluates the proposed method in terms of coverage rate.

1. Introduction

Inference for small domain parameters, such as average income for Blacks, or total number of people with body mass index (BMI) greater than 30 in a small state, under multistage complex sampling design is very challenging because of the limited sample size for the small domains. Traditionally, degrees of freedom (*df*) are approximated using the total number of primary sampling units (PSU) minus the total number of first stage strata to conduct inference for both national and domain parameters. However, such estimates may overestimate the true *df*, which has been discussed by Rust and Rao (1996) and Valliant and Rust (2010), among others. One improvement is to estimate *df* for small domain parameters by using a rule of thumb (RT), which is the total number of PSUs that have at least one element in the domain, minus the total number of strata that have at least one element in the small domain. This has been suggested by Rust and Rao (1996), Korn and Graubard (1999), and Burns et al. (2003). Burns et al. (2003) did some empirical studies and showed some benefits of the proposed method. Alternatively, Johnson and Rust (1993), Kott (1994), and Valliant and Rust (2010) proposed using Satterthwaite approximation, and the coverage rates are very close to the nominal rates if the distribution of variance estimator can be well described by the Chi-squared distribution.

The main motive for this research is relating to the limitation that most software packages do not adjust *df* for small domains automatically (Lewis, 2013); hence, some inconvenient manual adjustments need to be done. As an example, if using an online analytic systems (OAS), which produces estimates (e.g., for tables) in real-time, there is no chance for the user to adjust the *df*. In general, because of limited data sources or confidentiality concerns, users may not have access to microdata to help determine *df* and apply an RT method. Suppose we have 100 replicates in the dataset of OAS and the sample includes only 20 cases in a domain. In an OAS, the inference would be based on 99 *df* obtained from the traditional RT method for the domain because, without access to the number of PSUs and strata for the domain, the user could not adjust the *df*. Therefore, the automatic adjustments of *df* for submitted queries are in demand. In addition, most of the existing approaches for adjusting *df* are based on the Taylor linearization variance estimator. We evaluate the replication variance estimator in this paper with the following three research objectives:

1. To provide a better inference for parameters associated with small domains based on paired jackknife replication variance estimator (JK2), as proposed in Rust and Rao (1996) under multistage complex sampling designs with two PSUs per stratum.

2. To incorporate our proposed method into an OAS and provide an automatic adjustment of *df*.
3. To evaluate the proposed method through simulation by using 2011 National Health Interview Survey (NHIS) data.

This paper is organized as follows. In Section 2, we discuss rules of thumb as well as our proposed method for adjusting degrees of freedom for domain parameters. Some simulation results are presented in Section 3. We continue in Section 4 with some conclusions.

2. Rules of thumb for approximating degrees of freedom for a domain

In this section, we describe the traditional RT for approximating *df* for a domain *D*. Then an adjusted RT is described, as well as a proposed RT for the paired jackknife replication variance estimator.

2.1 Basic setups and traditional rule of thumb approaches

Suppose we have a finite population $(x_{hij}, y_{hij}, D_{hij}), h = 1,2, \dots, H, i = 1,2, \dots, M_h, j = 1,2, \dots, N_{hi}$, where H, M_h and N_{hi} are the number of strata, number of first stage units in stratum h and number of second stage units in PSU i which is in stratum h, respectively. Then, x_{hij},y_{hij}, and D_{hij} are the covariate, study variable, and domain indicator, respectively, for unit i such that $D_{hij} = 1$ if unit j belongs to domain *D* and $D_{hij} = 0$, otherwise. For simplicity, we only assume the two-stage stratified design with m_h PSUs selected in stratum h by using systematic probability proportional-to-size (PPS) sampling design with selection probability $\pi_{hi} = m_h N_{hi} / \sum_{i=1}^{M_h} N_{hi}$ and sampling weight $w_{hi} = \pi_{hi}^{-1}$. The second-stage sampling is assumed to be simple random sampling within each PSU to achieve self-weighting with overall target sample size n. In other words, the second-stage conditional inclusion probability is $\pi_{j|hi} = r/\pi_{hi}$, where $r = n/N$ is the target sampling rate. Suppose the parameter of interest is the domain mean of study variable $\bar{\theta}_D = N_D^{-1} \sum_{h=1}^{H} \sum_{i=1}^{M_h} \sum_{j=1}^{N_{hi}} D_{hij} y_{hij}$, with N_D as the population size for domain *D*. The traditional Hajek estimator (Hajek, 1971) can be written as

$$\hat{\bar{\theta}}_D = \frac{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} w_{hi} w_{j|hi} D_{hij} y_{hij}}{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} w_{hi} w_{j|hi} D_{hij}} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} D_{hij} y_{hij}}{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} D_{hij}}$$

where $w_{j|hi} = \pi_{j|hi}^{-1}$ is the second-stage conditional weight.

The traditional stratified jackknife variance estimator for two or more PSUs per stratum (JKn), discussed in Shao and Tu (1995) and Wolter (2007), can be written as:

$$\hat{V}_{JKn} = \sum_{h=1}^{H} \frac{m_h - 1}{m_h} \sum_{k=1}^{m_h} (\hat{\bar{\theta}}_{D(k)} - \hat{\bar{\theta}}_D)^2$$

where $\hat{\bar{\theta}}_{D(k)}$ is the estimate after deleting the kth PSU from stratum h. Specifically, we have

$$\hat{\bar{\theta}}_{D(k)} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} w_{hi}^{(-k)} w_{j|hi} D_{hij} y_{hij}}{\sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{n_{hi}} w_{hi}^{(-k)} w_{j|hi} D_{hij}}$$

with $w_{hi}^{(-k)} = m_h (m_h - 1)^{-1} w_{hi}$ if $k \in U_h$ and $i \neq k$, $w_{hi}^{(-k)} = 0$ if $k \in U_h$ and $i = k$ and $w_{hi}^{(-k)} = w_{hi}$, otherwise. In terms of *df* estimation for $\hat{V}_{JKn}$, the traditional RT has been used frequently, which estimates the *df* by the number of PSUs (m) – number of strata (H), where $m = \sum_{h=1}^{H} m_h$ (see Heeringa et al. (2010), among others). However, it is well known that the traditional RT overestimates the true *df* (see, for example, Rust and Rao (1996), Valliant and Rust (2010)).

2.2 Adjusted rule of thumb and Satterthwaite approaches

An alternative approach is to adjust the RT by estimating the *df* as the number of PSUs with at least one element in domain D (m_D) minus number of strata (H_D) with at least one element in domain D. The adjusted RT has been suggested in Rust and Rao (1996) and Korn and Graubard (1999) for inference with domain parameters. Burns, et al. (2003) conducted a simulation study and found that the *df* produced by the adjusted RT is much smaller than the traditional RT method for small domain estimation and the adjusted RT outperforms the traditional RT in terms of hypothesis testing.

For any replication method, Bryant (1994) suggested first producing multiple variance estimators based on several random subsets of the original replication weights, then estimating the *df* by using Satterthwaite approximation, which is $\widehat{df} = 2[\hat{E}\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\}]^2/\widehat{Var}\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\}$, where $\hat{E}\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}}\}$ and $\widehat{Var}\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\}$ are the sample mean and variance based on the multiple variance estimators created previously. In the paper, simulation studies show that the proposed methods are not stable in terms of estimating *df.* Prior to that, Johnson and Rust (1993) proposed several empirically derived effective *df* by using Satterthwaithe approximation and tested the effectiveness by using National Assessment of Educational Progress data. Kott (1994) proposed using Satterthwaite approximation of *df* for hypothesis testing of domain-based regression coefficients. For a single stage design, Valliant and Rust (2010) compared the traditional RT with the Satterthwaite method in a simulation study and found that the Satterthwaite approximation was closer to the true Monte Carlo *df* and the corresponding confidence interval had better coverage rates. The Satterthwaite method assumes $\hat{V}(\hat{\bar{\theta}}_D)$ follows the Chi-squared distribution with *df,* where $\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})$ denotes either Taylor linearization or replication variance estimators. Because $E\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\} = df$ and $Var\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\} = 2df$, then the df can be estimated by $\widehat{df} = 2\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\}^2/\widehat{Var}\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\}$, where $\widehat{Var}\{\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})\}$ is the variance estimator of $\hat{\mathrm{V}}(\hat{\bar{\theta}}_{\mathrm{D}})$. More details are presented in Valliant and Rust (2010). Most of the existing methods consider the Taylor linearization variance estimator except Johnson and Rust (1993) and Bryant (1994).

2.3 Adjusted rule of thumb for paired jackknife

In this section, we assume $m_h = 2$ for $h = 1,2, \ldots H$. We consider the paired jackknife estimator (JK2) proposed in Rust and Rao (1996). The JK2 approach produces a good balance between the number of replication weights and precision of the variance estimator. Basically, one of the two original replication weights from JKn is randomly selected for each stratum, which generates H replication weights using JK2. The variance estimator for JK2 can be written as

$$\hat{V}_{JK2} = \sum_{k=1}^{H}(\hat{\bar{\theta}}_{D(k)} - \hat{\bar{\theta}}_D)^2,$$

where $\hat{\bar{\theta}}_{D(k)}$ is the estimate after deleting the k^{th} PSU from stratum h and it is defined in Section 2.1.

The traditional RT for estimating *df* of $\hat{V}_{JK2}$ is the number of strata H. The *df* produced by the adjusted RT can be approximated by $H_{D,}$ which is the number of variance strata H with at least one element in D. In order to produce better inference for small domains, and to further reduce the number of replication weights, Nixon, et al. (1998) proposed a combined jackknife method (CJK2) to create replication weights. The following three steps describe the CJK2 procedure:

Step1: Sort the variance units by variance strata.
Step 2: Combine the adjacent two variance strata (e.g., combine variance stratum 1 with variance stratum 2, variance stratum 3 with variance stratum 4, and so on) and form two larger PSUs by randomly pairing one PSU from each variance stratum.
Step 3: Create paired jackknife replication weights based on the combined variance strata and variance units.

After creating the CJK2 replication weights, and by using the similar procedures as JK2, the corresponding traditional RT and adjusted RT can be computed. Table 1 in the Appendix summarizes the comparison among existing and proposed methods.

3. Simulation study

A simulation study conducted using the 2011 NHIS public use file with a sample size of 33,014 of sampled adults is discussed in this section. The Sample Adult dataset was downloaded from the Centers for Disease Control and Prevention website: http://www.cdc.gov/nchs/nhis/nhis_2011_data_release.htm (accessed 12/18/2013). After deleting cases with missing values for study variables AHEIGHT (Height), AWEIGHTP (Weight), and BMI (Body Mass Index), there were 30,075 cases remaining in the data set, which was treated as the finite population for simulation study. There were 1,000 Monte Carlo samples selected using a two-stage stratified PPS without replacement sampling design for each Monte Carlo sample. Thirty strata were created by combining STRAT_P (variance stratum) 1-10, 11-20…291-300. After combining strata, there were $M_h = 20$ PSUs per each stratum h. After creating the strata $m_h = 2$, PSUs were selected within each stratum h by using a systematic PPS design, where the size measure was the total population per PSU. Conditioning on the selected PSUs, persons were selected via simple random sampling without replacement (SRSWOR) in each PSU in order to produce an overall equal probability sample. The following two variance estimators were considered:

1. Paired Jackknife (JK2) variance estimator described in Section 2.3; and
2. Combined Strata Paired Jackknife (CJK2) variance estimator described in Section 2.3.

The *df* were estimated by the following two approaches for both JK2 and CJK2:

1. The traditional RT described in Section 2.3; and
2. The adjusted RT described in Section 2.3.

Three significance levels (0.01, 0.05, and 0.1) were considered to construct confidence intervals for the two study variables AWEIGHTP (Weight) and BMI (Body Mass Index). We considered two minority domains in the study, namely, Black (15% of total population) and Asian (6% of total population). We used $n = 200$ as the overall sample size. For JK2, there were 30 strata and 60 PSUs. For CJK2, there were 15 strata and 30 PSUs. The average sample sizes per PSU are presented in Table 2 in the Appendix. The following five parameters of interest were investigated:

1. Average BMI (Body Mass Index);
2. Average AWEIGHTP (Weight);
3. Regression coefficient (Slope) of AWEIGHTP (Weight) versus AHEIGHT (Height);
4. Total number of Black (or Asian) with BMI > 30 (or BMI > 25); and
5. Traditional RT estimation of *df* and adjusted RT estimation of *df* for all the above cases.

For parameters 1-4 above, we compared the two approaches based on the coverage rate, which is equal to the percentage of samples for which the confidence interval resulting from the Monte Carlo sample contains the true value of the parameter. Ideally, it should be equal to the nominal level (e.g., 95% for a significance level equal to 0.05). For parameter 5, we calculated the Monte Carlo bias, which is presented in Table 3. We found that both the traditional RT and adjusted RT approaches overestimate the true *df,* but the adjusted RT approach has a smaller Monte Carlo bias than the traditional RT. According to Figures 1-3 in the Appendix, for parameters 1-3 and Black or Asian domains, the adjusted RT approach has better coverage rates than the traditional RT approach. In other words, the differences between the simulation coverage rates and target coverage rates are smaller. The intervals for the Black domain are much closer to the nominal coverage rate than that for Asians since there are fewer Asians in the population. All the confidence intervals have smaller coverage rates than the nominal rates, which means that the estimators for *df* have positive bias. According to Figure 4, both traditional RT and adjusted RT approaches for the "Total number of Asians with BMI>30" produce unacceptable coverage rates, because of the extremely rare corresponding domain. As seen in Table 4, the expected sample size for "Asians with BMI>30" is very small, which may contribute to the inadequate performance in terms of coverage rates. According to Figure 5, for parameter 4 "Total number of Asians with BMI>25," the coverage rates are much better than previous parameter "Total number of Asians with BMI>30," which confirms that the sample size for the domain makes a large impact on the estimation of *df*. The results are constrained by the simulation setup. Therefore, we note that the evaluation was focused on a stratified multistage PPS with two PSUs per stratum for the *df* adjustment approach. They are further limited by the data set considered, which was the NHIS with study variables AHEIGHT (Height), AWEIGHTP (Weight) and BMI (Body Mass Index). The parameters of interest included only means, regression coefficients, and

totals. In our simulation study; the sample selection for the simulation is assumed to be spread evenly across PSUs within strata.

4. Conclusions

The adjusted RT approach for inference of domain parameters through an automated adjustment of *df* while using a version of the Paired Jackknife (JK2 or CJK2) has been evaluated through a simulation study using 2011 NHIS data. Under the simulation setup, there was some improvement seen in the adjusted RT approach over traditional RT for means, regression coefficients, and totals for larger domains in terms of coverage rates and the estimation of *df.* For a total for a very small domain (< 1%), neither the traditional nor adjusted RT approaches performed well. The benefits of the adjustment are greatest when there are few PSUs, e.g., when producing estimates for geographical domains. Furthermore, the adjusted RT method can be considered as having potential improvement over the traditional RT approach under a potential application to OAS or other software that produces results in real-time.

The following items would be interesting to pursue in future research:

- Consider balanced repeated replication (BRR) and Fay's BRR for estimating means, totals, regression coefficients, and even quantiles or quantile regression.
- Compare the traditional stratified jackknife variance estimator with adjusted RT Satterthwaite's method, and Bryant (1994)'s method with proposed methods in a unified way.
- Estimate *df* for the estimates that incorporate information from multiple surveys. For example, the control totals for weight calibration can be obtained from another complex survey. In this case, we may need to use the replicate weights from both surveys to determine the *df* for the calibrated estimates.
- Create multiple bootstrap variance estimators based on Booth, Butler, and Hall (1994) and Chauvet (2007). Then use Satterthwaite's approximation to estimate *df.*

Acknowledgements

The authors are very grateful to Keith Rust and Graham Kalton for their valuable comments.

References

Bryant, E. (1994). Methodological issues in comparative educational studies: *The case of the IEA Reading Literacy Study. Chapter 2*. *U.S. Department of Education, National Center for Education Statistics: Washington, D.C.*

Burns, A. M., Morris, R. J., Liu, J., and Byron, M, Z. (2003). Estimating degrees of freedom for data from complex surveys. *2003 Joint Statistical Meetings - Section on Survey Research Methods*.

Booth, J. G., Butler, R. W. and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association.* **89**, 1282–1289.

Chauvet, G. (2007). Methodes de bootstrap en population finie. PhD Thesis, Universite de Rennes 2.

Hajek, J. (1971). Comment on a paper by D. Basu. In: VP. Godambe and D.A. Sprott (eds.), *Foundations of Statistical Inference*. Holt, Rinehart and Winston, Toronto, Ontario, Canada, p. 236.

Heeringa, S. G., West, B. T., and Berglund, P. A. (2010). *Applied Survey Data Analysis*. Chapman & Hall / CRC Press, Boca Raton, FL.

Johnson, E. G. and Rust, K. (1993). Effective degrees of freedom for variance estimates from a complex sample survey. Unpublished manuscript, presented at the 1993 Joint Statistical Meetings.

Kott, P. S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology* **20**, 159-164.

Korn, E. L., and Graubard, B. I. (1999). *Analysis of health surveys*. New York: Wiley.

Lewis, T. (2013). Considerations and techniques for analyzing domains of complex survey data. *SAS Global Forum, Statistics and Data Analysis.*

Nixon, M. G., Brick, J. M., Kalton, G., and Lee, H. (1998). Alternative variance estimation methods for the NHIS. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Rust, K. F., and Rao, J. N. K. (1996). Variance estimation for complex surveys using replication. *Statistical Methods in Medical Research*, **5**, 283–310.

Shao, J., and Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer-Verlag.

Valliant, R., and Rust, K. (2010). Degrees of freedom approximations and rules-of- thumb. *Journal of Official Statistics*, **26**, 585-602.

Wolter, K. M. (2007). *Introduction to variance estimation*, 2nd ed. New York: Springer-Verlag.

Appendix: Tables and Figures

Table 1 Comparison of different approaches for estimating the *df*

Select literature	*Df*	Variance approach	Estimates	Inference
Johnson and Rust (1993)	Satterthwaite	General replication	Mean, *df*	point estimate, standard error (SE), confidence interval (CI)
Bryant (1994)	Satterthwaite	General replication	*df*	point estimate, SE
Kott (1994)	Satterthwaite	Taylor	regression coefficient	hypothesis testing
Korn and Graubard (1999)	Adjusted RT	Taylor	mean	Mean, SE
Burns et al. (2003)	Adjusted RT	Taylor	mean, *df*	point estimate, hypothesis testing
Valliant and Rust (2010)	Traditional RT, Satterthwaite	Taylor	total, ratio, *df*	point estimate, SE, CI
This paper	Adjusted RT	JK2, CJK2	mean, regression coefficient, total, *df*	point estimate, SE, CI

Table 2 Average sample sizes per PSU

Domain	JK2	CJK2
Overall	3.3	6.6
Black	0.5	1.0
Asian	0.2	0.4

Table 3 Monte Carlo bias for the estimated degrees of freedom

Approach	Domain	Diff between est *df* and target *df*	Parameter				
			Avg. Height	Avg. Weight	Avg.BMI	Regression Coefficient	Total BMI>30
JK2	Black	Traditional	20.0	21.3	22.0	25.2	18.6
		Adjusted	6.2	7.5	8.2	11.4	4.8
	Asian	Traditional	25.5	28.1	27.7	29.3	27.9
		Adjusted	4.4	7.0	6.5	8.1	6.7
CJK2	Black	Traditional	7.8	8.1	9.0	11.2	7.4
		Adjusted	4.3	4.6	5.5	7.7	3.9
	Asian	Traditional	11.2	12.7	13.3	14.8	13.0
		Adjusted	3.4	4.8	5.4	7.0	5.2

Table 4 Expected sample size in each domain

	Black	Asian	Overall
Total	30	13	200
BMI>25	22	5	126
BMI>30	11	1	55

Figure 1 Simulation coverage rates – average BMI

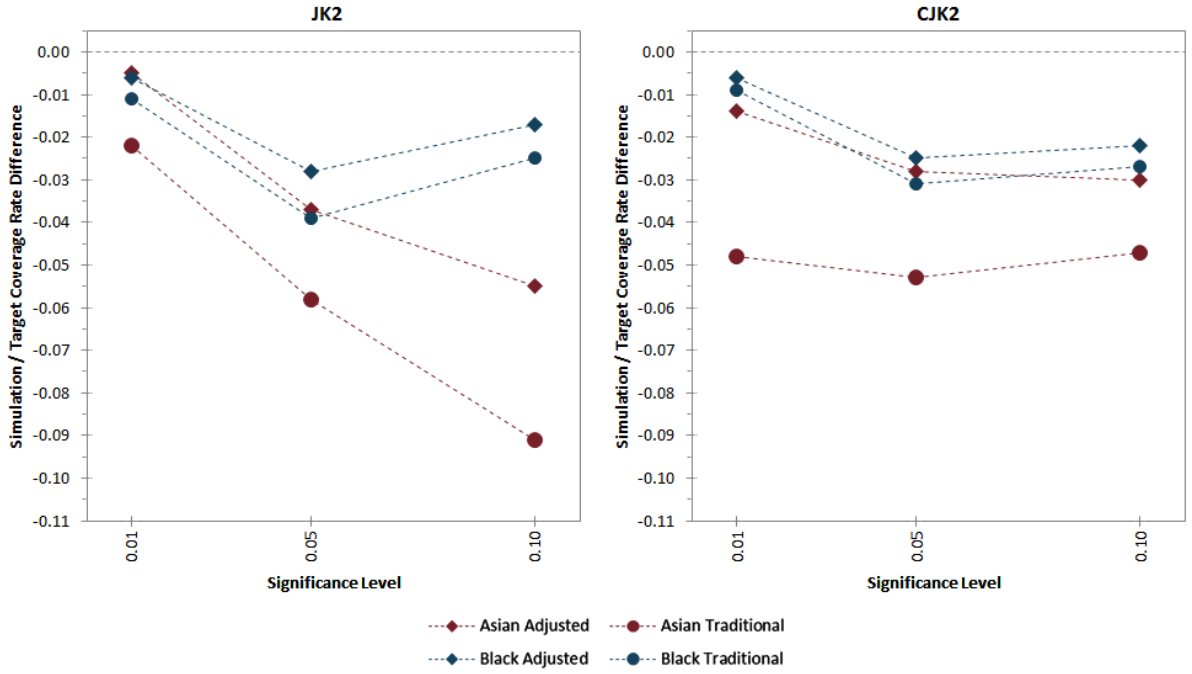

Figure 2 Simulation coverage rates – average weight

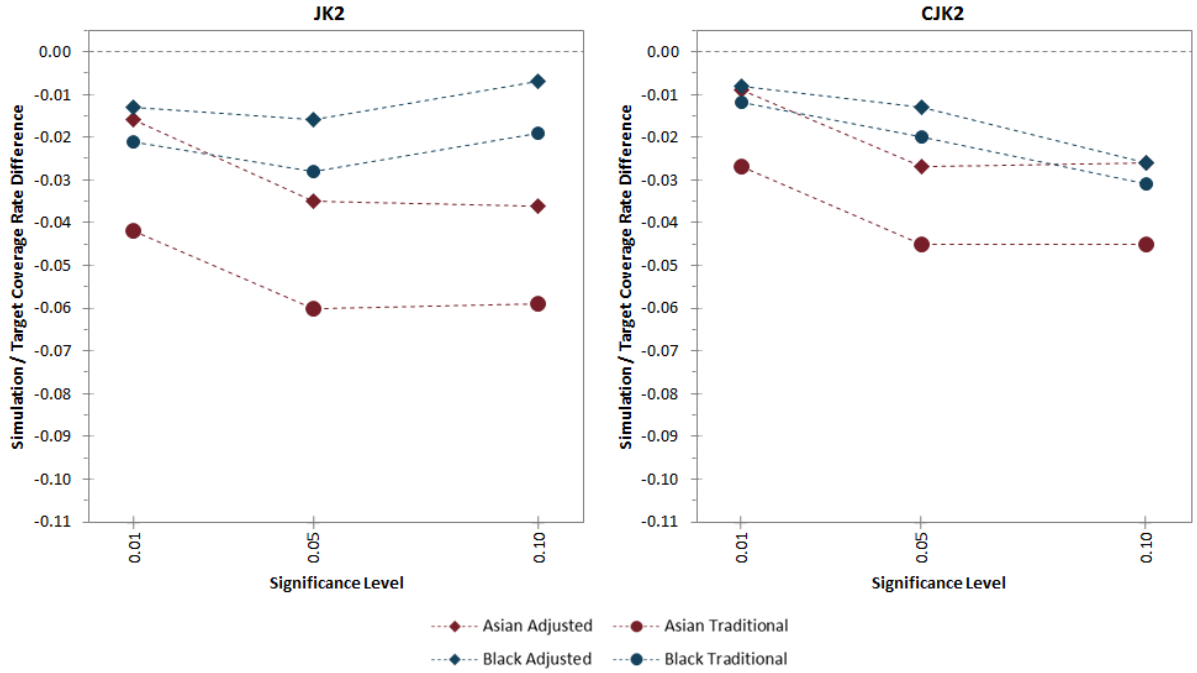

Figure 3 Simulation coverage rates – regression coefficient

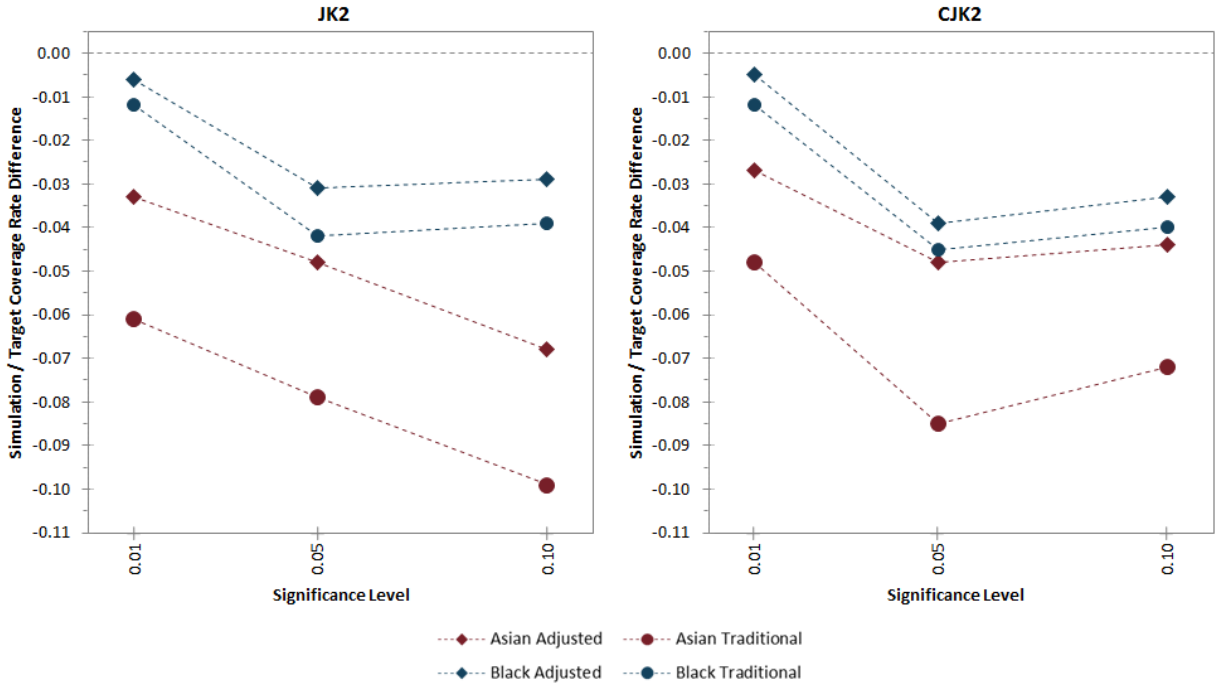

Figure 4 Simulation coverage rates – total with BMI > 30

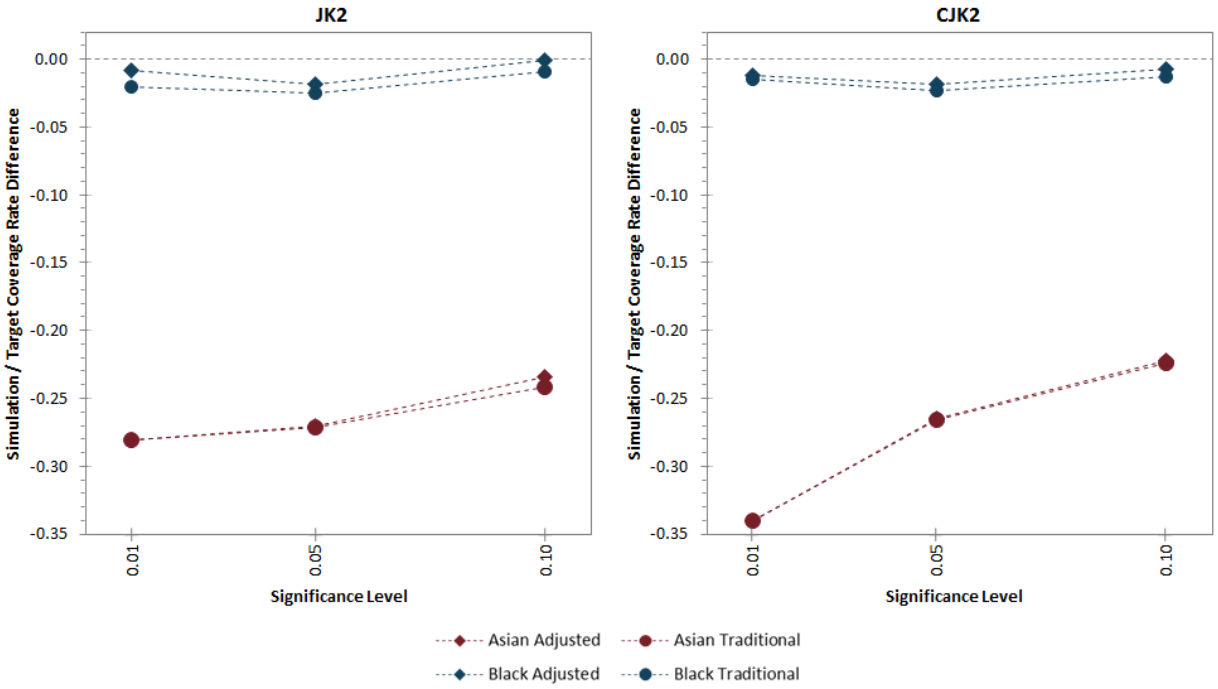

Figure 5 Simulation coverage rates – totals for Asians with BMI > 30 and BMI > 25