ABSTRACT

Standardized errors \( \frac{\hat{T} - T}{\nu} \sqrt{\frac{1}{2}} \) were calculated for both ratio and regression estimators in each of 10,000 simple random samples of \( n = 32 \) from each of six populations, using four different variance estimators. Graphs show how the percentage of intervals \( \hat{T} \pm 1.96 \sqrt{\nu} \) which fail to contain \( T \) changes as a function of the average value of the auxiliary variable in the sample. They reveal that (i) Intervals using the variance estimators from standard linear regression theory were hopelessly unreliable; (ii) Intervals using the conventional finite population variance estimators showed a striking excess of failures in badly balanced samples, and (iii) None of the four variance estimators produced satisfactory confidence intervals in populations arising from badly skewed prediction models.

1. INTRODUCTION

Previous empirical studies (Royall and Cumberland 1981a, b) confirmed prediction theory’s warnings about bias in ratio and regression estimators of population totals when the sample is badly balanced, about failure of conventional variance statistics to estimate properly conditioned variances, and about non-robustness of linear regression theory’s variance estimators. Those studies showed the importance of good balance for robust estimation and the value of variance estimators which prediction theory had identified as bias-robust.

In discussing the ratio estimation paper, T.M.F. Smith (1981) commented that “My only criticism ... is that we will have to wait to discover what are the coverage properties of the intervals based on the various variance estimators.” Here we report results of a large scale empirical study of coverage properties of confidence intervals about ratio and regression estimators, using the standard normal approximation with various estimates of standard error, in samples drawn at random from the same study populations used in the previous papers.

The results will be stated in terms of the standardized error \( \text{SZE}, \frac{\hat{T} - T}{\nu} \sqrt{\frac{1}{2}} \), where \( \hat{T} \) is either the ratio \( \hat{T}_1 \) or regression \( \hat{T}_2 \) estimator of the population total \( T \), and \( \nu \) is a variance estimator. The variance estimators for \( \hat{T}_1 \) which were denoted in the previous study (1981a) by \( \nu_C, \nu_D, \nu_J, \) and \( \nu_L \) appear here as \( \nu_{1C}, \nu_{1D}, \nu_J, \) etc., and the corresponding estimators for the variance of \( \hat{T}_2 \) (1981b) appear as \( \nu_{2C}, \nu_{2D}, \) etc. Whenever \( |\text{SZE}| > k \) the corresponding confidence interval \( \hat{T} \pm k \sqrt{\nu} \sqrt{\frac{1}{2}} \) fails to include the true total \( T \).

The previous work (Royall and Cumberland 1981a, b) led to conjectures about the behavior of the SZE’s in the six study populations. For example, in those populations where an estimator \( T \) shows a bias, we expect that when the bias is positive the frequency of SZE > k will exceed that of SZE < -k, with the opposite occurring when \( T \) has a negative bias. The discrepancy will be greatest when the bias is strongest.

Other conjectures concern the effects of biases in the variance estimators. In those populations with no clear bias in \( \hat{T} \), we expect that

(i) With the conventional variance estimators the absolute standardized errors \( \left| \frac{\hat{T}_1 - T}{\nu_{1C}} \right| \sqrt{\frac{1}{2}} \) and \( \left| \frac{\hat{T}_2 - T}{\nu_{2C}} \right| \sqrt{\frac{1}{2}} \) will exceed a positive constant \( k \) more frequently than predicted by the (normal or Student's t) reference distribution in the samples where \( \bar{x}_s < \bar{x} \), and the frequency will decrease as \( \bar{x}_s \) increases, becoming less than the reference frequency when \( \bar{x}_s \) is much greater than \( \bar{x} \).

(ii) For \( \hat{T}_2 \) and \( \nu_{2L} \) performance should be qualitatively similar to that for \( \nu_{2C} \) in (i), but with somewhat lower frequencies when \( |\bar{x}_s - \bar{x}| \) is large.

(iii) For \( \hat{T}_1 \) and \( \nu_{1L} \) the frequency should exceed that of the reference distribution uniforming in \( \bar{x}_s \).

(iv) Replacing \( \nu_{1C} \) and \( \nu_{1L} \) \( (j = 1 \) or \( 2 \)) by \( \nu_{1D} \) or \( \nu_{1J} \) should improve the approximation to the reference distribution.

These conjectures come directly from the theoretical and empirical results on expected values of the various statistics. Consider for example, the variance estimators \( \nu_{1C} \) and \( \nu_{1L} \).

The fact that in samples where \( \bar{x}_s < \bar{x} \) the average value of \( \nu_{1C} \) was smaller than the average of \( (\hat{T}_1 - T)^2 \) leads us to expect an excess of extreme values of \( \left| \frac{\hat{T}_1 - T}{\nu_{1C}} \right| \sqrt{\frac{1}{2}} \) in such samples. The fact that the average value of \( \nu_{1L} \) was smaller average \( (\hat{T}_1 - T) \) over the
entire range of values of $\bar{x}_s$ leads us to expect an excess of extreme values of $|\hat{T}_1 - T|/\sqrt{\hat{v}_{1L}}$ everywhere.

A major source of uncertainty in these conjectures is the effect of non-normality. The $Y$-variables studied in finite populations are often non-negative and sharply skewed, so that $\hat{T} - T$ does not have an approximate normal distribution until $n$ becomes very large. The skewness can also distort the distribution of $v$ and produce a correlation between $T$ and $v$, sharply altering the distribution of $SZE$.

2. EMPIRICAL STUDY

We examined the behavior of $SZE$'s using $\hat{T}_1$, $\hat{T}_2$, and various variance estimators in the same six populations used in the previous studies. Descriptive statistics as well as scatter diagrams for all populations appear in Royall and Cumberland (1981a). From each population we drew ten thousand simple random samples of $n = 32$, and for each sample we calculated the $SZE$, $(\hat{T} - T)/\sqrt{v}$, for $\hat{T} = \hat{T}_1$ with $v = v_{1L}$, $v_{1C}$, $v_{1D}$, and $v_{1J}$, and for $\hat{T} = \hat{T}_2$ with $v = v_{2L}$, $v_{2C}$, $v_{2D}$, and $v_{2J}$. To see how performance varied with $\bar{x}_s$, we ordered the ten thousand samples according to their values of $\bar{x}_s$, and divided them into twenty groups of five hundred samples, the first group containing the five hundred samples whose $\bar{x}_s$ are smallest, etc. For each group and for each combination $(\hat{T}, v)$ studied we calculated the percentage of the five hundred $SZE$'s exceeding 1.96 and the percentage falling below $-1.96$. Figures 1a and 1b show these percentages for the ratio estimator $\hat{T}_1$ for each of the six populations. Figures 2a and 2b show the corresponding percentages for the regression estimator $\hat{T}_2$.

3. RESULTS

If the standard normal approximation to the distribution of a $SZE$ is to provide useful confidence intervals then on the corresponding graph each of the twenty bars should be about five percentage units long, and each should be centered near zero. A slightly more conservative approximation replaces the normal by the Student's $t$ distribution with 31 d.f. for the ratio and 30 d.f. for the regression estimator. With these $t$ distributions the bars should be about six percentage units long. However, the graphs show that it is unrealistic to fret about a couple of percentage points or degrees of freedom in the situations represented here. The problems are much more serious.

The results show that the prediction theoretical results concerning bias in $\hat{T}_1$ and $\hat{T}_2$, expected values of the variance estimators, and the importance of balanced samples in robust inference are directly relevant to confidence intervals. In Figure 1a the Cancer population provides a comparison of the variance estimators' performance in the absence of serious bias in $\hat{T}_1$. The negative bias in $v_{1L}$ in this population translates directly into an excess of extreme $SZE$'s throughout the range of $\bar{x}_s$, with over ten percent of $|SZE|'$s exceeding 1.96 in every group. The conventional variance estimator's increasing bias, from negative when $\bar{x}_s$ is small to positive when $\bar{x}_s$ is large, produced the funnel shaped graph showing that as $\bar{x}_s$ increased the percentage of extreme values of $|\hat{T}_1 - T|/\sqrt{v_{1C}}$ decreased steadily from 27 to 3. The bias-robust statistics $v_{1D}$ and $v_{1J}$, although producing a slight excess of extreme values in every group, showed much better performance than either $v_{1L}$ or $v_{1C}$. The Sales population produced similar results.

In Counties 60 (Figure 1a) and Hospitals (Figure 1b) the effects of failure of the simple proportional regression model appear. The previous study (Royall and Cumberland 1981a, Figures 9 and 11) showed that in both of these populations the average value of $\hat{T}_1 - T$ is large (relative to the variability) and positive when $\bar{x}_s$ is small, drops to zero when $\bar{x}_s$ is near $x$, and becomes large and negative when $\bar{x}_s$ is large. The apparent cause in both populations is a regression function, $E(Y)$, which increases with $x$, but with a steadily decreasing slope. In the Counties 60 population this produced striking results (Figure 1a). Most remarkable is the disastrous performance of the conventional variance esti-
mator in the group of five hundred samples

whose $\bar{x}_s$ values are smallest, where fully

65% of the standardized errors exceeded 1.96.
(In this group nearly 23% of the SZE's exceeded

4.0) The standard regression variance estima-
tor, $v_{1L}$, produced an excess of extreme SZE's
in every group, with a gross excess at both
extremes of the $\bar{x}_s$ distribution. The estimators

$v_{1D}$ and $v_{1J}$ also performed badly at the
extremes, but unlike $v_{1L}$ both they and $v_{1C}$ gave
acceptable, though conservative, results in
well-balanced samples. The Hospitals popula-
tion (Figure 1b) produced similar graphs,
though less dramatic than Counties 60. We
interpret these results as showing that balance is a prerequisite for robust inference using
the ratio estimator and any of these or similar
variance estimators. In the 10-20 percent of
randomly selected samples showing the worst
balance, nominal 95% confidence intervals can
have coverage probabilities much lower than

0.95.

The last two populations, Counties 70 and
Cities (Figure 1b), are the most alarming. In
Counties 70 many relationships remained the same
as in the other populations: $v_{1L}$ gave too many
extreme SZE's in every group, $v_{1C}$ gave the fun-
nel shape expected, and $v_{1D}$ and $v_{1J}$ provided
more stable and consistent performance, over the
twenty groups, than the others did. But
these two bias-resistant statistics also produced a striking excess of large SZE's in every
group, even those where $\bar{x}_s$ was near $\bar{x}$. This
is curious, because the earlier study showed that
for this population the average error,

$\hat{T}_1 - T$, is nearly zero in well-balanced
samples and is positive when $\bar{x}_s$ is greater
than $\bar{x}$ (Royall and Cumberland 1981a, Figure 10).

Thus the analysis in terms of biases in $\hat{T}_1$

(caused by failure of the proportional regression model), which seemed to explain the
excess of extreme SZE's in the badly balanced
groups of samples from Counties 60 and
Hospitals, does not apply here. One possible
explanation for the Counties 70 results is to be
found in the positive covariance between

$\hat{T}_1$ and $v_{1/2}$, which produced a correlation co-
efficient between 0.76 and 0.82 within every
one of the twenty sets of samples for both

$v_{1D}$ and $v_{1J}$. Thus when $\hat{T}_1$ is small, giving a
negative error, $\hat{T}_1 - T$, the variance estimators
also tend to be small, producing large negative
standardized errors. On the other hand, when
when $\hat{T}_1 - T$ is positive the variance estimators
tend to be large, preventing a large positive
standardized error. The source of these corre-
lations is uncertain, but they are probably
cased by skewness in the Y-distribution. The
same phenomenon appears to have occurred in the
Counties population (correlation coefficients
between 0.70 and 0.76), although it is not
nearly so severe in that case.

Figures 2a and 2b show that for the
regression estimator also the earlier results
on how the average values of $T_2$, $(T_2 - T)^2$, and
the various variance estimators relate to $\bar{x}_s$
to the SZE's very well in these
populations, again with the exception of
Counties 70 and possibly Counties. Again the
tendency of $v_{2C}$ to increase sharply with in-
creasing $\bar{x}_s$ produced funnel shaped graphs indi-
cating a gross excess of extreme SZE's when
$\bar{x}_s$ is small. Again $v_{2D}$ performed much better,
better, as did $v_{2J}$. And as before, even the
best variance estimators produced grossly unsatisfactory results in the Counties 70 popu-
lation where again a large positive correlation between the numerator and denominator of SZE
was observed. We note that in these popula-
tions the jackknife statistic $v_{2J}$ performed
noticeably better than $v_{2D}$, although the dif-
fences were not nearly as great as those
between these two statistics and the other two,
$v_{2L}$ and $v_{2C}$.

The qualitative results are insensitive to
the value 1.96 chosen as the reference point in the figures. For example, in the first group
(smallest $\bar{x}_s$ chosen as the reference point in the figures). For example, in the first group
(population with $\bar{x}_1$ and $v_{C}$ the percentage of
SZE's $< -1.96$ is 45, much greater than the
nominal value 2.5. If $k = -1.96$ is replaced by
$k = -1.64$, -2.58, and -4 respectively, the
the corresponding percentages are 49, 38, and 21.
The variance estimators from standard regression theory, $V_{1L}$ and $V_{2L}$, are so non-robust to changes in the form of $\text{var}(Y)$ (as a function of $x$) that they are useless in practice. This suggests the possibility that outside of finite populations, the standard formulas for estimating the variance of a regression coefficient should be replaced in routine practice by more bias-robust alternatives.

The flaws which prediction models revealed in the current favorites, $V_{1C}$ and $V_{2C}$, are fatal. These statistics should be replaced by more bias-robust alternatives such as $V_{1D}$ and $V_{2D}$ or $V_{1J}$ and $V_{2J}$.

Balance on $x$ is a necessary condition for robust inference using the ratio estimator, as the results for Counties 60 and Hospitals show. But the Counties 70 graphs warn that balance is not sufficient for robust interval estimation using the standard normal approximation, even with the best of the four variance estimators studied here. We are studying other interval estimation procedures in hopes of finding, for populations like Counties 70, confidence intervals which will live up to their name.

A popular belief among sampling statisticians is that inferences based only on the random sampling distribution are robust. These results show that their faith is misplaced. Clearly, the sampler who sets confidence intervals using either the ratio or regression estimator with its conventional variance estimator and who feels that his inferences are robust because the sample was chosen at random, is wrong. Inferences must be made conditionally on observable characteristics of the sample drawn, and they require assumptions regarding the population structure which are most naturally expressed through prediction models.

REFERENCES


Figure 1a. Standardized Error of Ratio Estimator with $v_{1L}$, $v_{1C}$, $v_{1D}$, $v_{1J}$:
Percentage of Extreme Values in Cancer, Sales, and Counties 60 Populations.a

a10,000 Simple Random Samples of $n = 32$ on $\bar{x}_n$ and divided into 20 groups of 500 samples. Reference lines are shown at 2.5 percent.
Figure 1b. Standardized Error of Ratio Estimator with $V_{LL}, V_{LC}, V_{LD}, V_{LJ}$:
Percentage of Extreme Values in Hospitals, Counties 70, and Cities Populations.
Figure 2a. Standardized Error of Regression Estimator with $v_{2L}$, $v_{2C}$, $v_{2B}$, $v_{2S}$: Percentage of Extreme Values in Cancer, Sales, and Counties 60 Populations.
Each bar represents one group of 500 samples. Groups are ordered on $\bar{y}_k$.

Figure 2b. Standardized Error of Regression Estimator with $v_{2L}$, $v_{2C}$, $v_{2D}$, $v_{2J}$: Percentage of Extreme Values in Hospitals, Counties 70, and Cities Populations.