Journal of Official Statistics, Vol. 17, No. 4, 2001, pp. 521–526

The Delete-a-Group Jackknife

Phillip S. Kott[1]

The National Agricultural Statistics Service of the U.S. Department of Agriculture has been using the delete-a-group jackknife (DAGJK) in an increasing number of its surveys. This article discusses the theory behind the DAGJK when the first-phase of sampling is stratified and there are a large number of sampled units per stratum, which is the case for many list-based surveys. It goes on to propose an extension of the DAGJK for use when the number of sampled units per stratum is less than the number of jackknife replicates.

Key words: Extended delete-a-group jackknife; replicate; stratum.

1. Introduction

The delete-a-group jackknife (DAGJK) is a relatively new name for a widely used procedure in survey sampling. For example, it is similar to ''Jackknife 1'' computed by WesVar (see Westat 2000, p. A-9). When done correctly, the DAGJK, like the conventional stratified jackknife (Rust 1985), can produce nearly unbiased estimates of mean squared error for a remarkably broad range of estimation strategies including many involving adjustments for nonresponse, calibration, composite estimation, and multi-phase sampling.

There are no theoretical advantages in using the delete-a-group rather than the stratified jackknife. Nevertheless, in list-based surveys where there are often thousands of sampled units in strata of varying sizes, the DAGJK offers computational advantages over its cousin. The DAGJK is simple to implement and easy to explain to external users of survey data.

Most establishment surveys are based on samples chosen randomly from a frame listing the members of the population of interest. In contrast, most demographic surveys in the US are based on multi-stage samples where a contiguous area, like a county, is selected at the first stage of sample selection. It is not uncommon for both list and area samples to be stratified, but with the latter there are often as few as two selections per stratum. Moreover, the first-stage of the sample design tends to be identical or nearly so across the strata.

With list-based surveys, the number of sample selections per stratum can vary widely. It is for this kind of design that the DAGJK has been developed. It is well known, however, that an expansion estimator based on a stratified simple random sample has a simple variance estimator no matter how varied the strata are in their sample sizes. The DAGJK is needed primarily when stratified random sampling is only the first of potentially several

[1] National Agricultural Statistics Service, Research Division, 3251 Old Lee Highway, Fairfax, VA 22050, U.S.A. E-mail: pkott@nass.usda.gov

phases of sampling or when the estimator of interest is more complex than an expansion estimator, for example, a regression coefficient.

One requirement for the appropriate use of the DAGJK is that the number of sample units per first-phase stratum be large in all strata. This is the situation in many, but not all, list-based sample surveys. An *extended DAGJK* is developed to treat situations where it is not. This formulation can be especially useful when a list-based and an area-based survey, the latter featuring few primary sample units per stratum, are combined to form multiple-frame estimates. Even in pure list-based surveys with hundreds of strata, it is not uncommon for at least some strata to have small sample sizes.

In brief, the DAGJK procedure divides the (first-phase) sample into R random groups and then estimates variances (or mean squared errors) by

1. deleting one group at a time from the sample,
2. computing R ''replicate'' estimates in an appropriate manner, and
3. taking the sum of the squared differences between the R replicate estimates and the original estimate multiplied by $(R-1)/R$.

Given a weighted estimator of the form $t = \sum_S w_k y_k$, say, the replicate-r ($r = 1, \ldots, R$) estimate has the form $t_{(r)} = \sum_S w_{k(r)} y_k$. The DAGJK variance estimator is

$$\text{var}(t) = ([R-1]/R) \sum^{R} (t_{(r)} - t)^2 \tag{1}$$

The key to this variance estimator is the development of the replicate-r weights, the $w_{k(r)}$. When the element k is within a primary sample unit (or is the sample unit itself), and the unit is a member of group r, $w_{k(r)}$ is set equal to zero. Otherwise, $w_{k(r)}$ is calculated by first adjusting the remaining w_k to account for those $w_{k(r)}$ that were set equal to zero.

In this article, we will restrict our attention to the common special case where the (first-phase) sample is stratified. In constructing a DAGJK, the primary sample units (PSUs) are first arranged with randomly-ordered units in the same stratum listed contiguously. From this ordering, the PSUs are systematically allocated into the R groups. In many list-based samples where the DAGJK will be of most use, the PSUs will be identical to the sample elements.

Suppose there are n_h PSUs in the same stratum (h) as the PSU containing element k, and n_{hr} PSUs in both the stratum and group r. Then a nonzero $w_{k(r)}$ is initially set to $[n_h/(n_h - n_{hr})]w_k$. This has not always been the rule in general practice, but it is what the theory in the next section dictates. The documentation for WesVar cited above goes so far as saying that jackknife 1 should not be used with a stratified sample, which is not the position taken here.

Other modifications may be necessary if w_k has been adjusted, say, for nonresponse or to match known population totals for a vector of auxiliaries. See Kott (1998) for the details, which are beyond the scope of this endeavor.

Section 2 shows why the DAGJK is reasonable for a simple expansion estimator under a single-phase, stratified sample when, one, finite population correction factors can be ignored, a common requirement with jackknives, and two, all stratum sample sizes and R are large. The DAGJK extends to smooth functions of expansion estimators and to

reweighted estimators based on multi-phase samples for reasons analogous to those for the stratified jackknife. See Rust (1985) and Kott and Stukel (1997), respectively, for the original arguments.

The near unbiasedness of DAGJK requires that the number of first-phase PSUs in each stratum be large. Kott (1998) puts the minimum number at five using the reasoning repeated at the end of Section 2. This requirement is not always met in practice even in list-based surveys. The resultant upward bias in the variance estimator may be acceptable in some situations. For others, the extended DAGJK is developed in Section 3.

Section 4 contains a brief discussion addressing why the National Agricultural Statistics Service has chosen to implement the DAGJK together with some recommendations about statistical testing and confidence-interval construction.

2. Justifying the DAGJK Under a Single-Phase, Stratified Sampling Design

Suppose we have a probability sample design with H strata and n_h PSUs within each stratum h. Let us assume that the sample was selected without replacement but the selection probabilities are all so small, and the joint selection probabilities are such, that using the with-replacement variance estimator is appropriate (this rules out systematic sampling from a purposefully-ordered list). In particular, let us assume that the estimator itself can be written in the form:

$$t = \sum_{h=1}^{H} \sum_{j=1}^{n_h} t_{hj}$$

where each t_{hj} is the sum of the $w_k\, y_k$ across the elements in PSU j of stratum h (which may contain only a single element in practice), and the w_k are the inverses of the element selection probabilities. Recall that extensions to more complex estimators and multi-phase samples (including quasi-designs that adjust for nonresponse), analogous to those for the stratified jackknife, are possible but beyond the scope of this article.

Let $q_{hj} = t_{hj} - t_{h+}$, where $t_{h+} = \sum t_{hg}/n_h$, and the summation is over the PSUs in h. The randomization variance of t is $\mathrm{Var}(t) = \sum^{H} \mathrm{Var}(t_{h+})$. Now $\mathrm{Var}(t_{h+})$ can be estimated in an (almost) unbiased fashion by

$$\mathrm{var}(t_{h+}) = (n_h/[n_h - 1]) \sum_{j=1}^{n_h} q_{hj}^2$$

(''almost'' because we are ignoring finite population correction).

In order to estimate $\mathrm{Var}(t)$ with a DAGJK, we first order the strata in some fashion and then order the PSUs within each stratum randomly. The sample is partitioned into R systematic samples using the resulting ordered list. Let S_r denote one such systematic sample, S_{hr} the set of n_{hr} PSUs in both S_r and stratum h, and $S_{h(r)}$ the set of $n_{h(r)}$ PSUs in stratum h and **not** in r.

The DAGJK replicate estimator $t_{(r)}$ is

$$t_{(r)} = \sum_{h=1}^{H} (n_h/n_{n(r)}) \sum_{j \in S_{h(r)}} t_{hj}$$

Now

$$t_{(r)} - t = \sum_{h=1}^{H} \left[(n_h/n_{h(r)}) \sum_{j \in S_{h(r)}} t_{hj} - t_{h+} \right]$$

Treating each $S_{h(r)}$ as a simple random subsample of the sample in stratum h, and taking expectations with respect to the subsampling with the sample fixed, we have

$$\begin{aligned} E_2[(t_{(r)} - t)^2] &= \sum^{H} \mathrm{Var}_2 \left([n_h/n_{h(r)}] \sum_{j \in S_{h(r)}} t_{hj} \right) \\ &= \sum^{H} (n_h^2/n_{h(r)})[1 - (n_{h(r)}/n_h)] \sum^{n_h} q_{hj}^2/(n_h - 1) \\ &= \sum^{H} (n_h/[n_h - 1])(n_{hr}/n_{h(r)}) \sum^{n_h} q_{hj}^2 \\ &= \sum^{H} (n_{hr}/n_{h(r)}) \mathrm{var}(t_{h+}) \end{aligned}$$

Observe that for strata where $n_h < R$, $n_{hr}/n_{h(r)}$ is either zero because there are no PSUs in both r and h or $n_{hr}/n_{h(r)}$ is $1/(n_h - 1)$ because there is one PSU in both r and h. Since the latter situation occurs in exactly n_h replicates, $\sum^R n_{hr}/n_{h(r)} = n_h/(n_h - 1)$.

For strata where $n_h \geq R$, $n_{hr}/n_{h(r)} = O(1/R)$ and $\sum^R n_{hr}/n_{h(r)} \approx 1 + O(1/R)$. (Technical note: $z = O(1/R)$ means $\lim_{R \to \infty} R|z|$ is a constant.) In fact, when n_h/R is an integer, $n_{hr}/n_{h(r)}$ exactly equals $1/(R - 1)$, and $\sum^R n_{hr}/n_{h(r)} = R/[R - 1]$.

Since $\mathrm{Var}(t)$ can itself be estimated in an approximately unbiased fashion by $\mathrm{var}(t) = \sum^H (n_h/[n_h - 1]) \sum_j q_{hj}^2$, it is not difficult to see that the DAGJK variance estimator in equation (1), $v_J = ([R - 1]/R) \sum^R (t_{(r)} - t)^2$, is approximately equal to $\mathrm{var}(t)$ – and thus approximately unbiased for $\mathrm{Var}(t)$ – when all strata are such that $n_h \geq R$ and is biased upward otherwise.

The relative upward bias in v_J is bounded by $([R - 1]/R) \max_h \{1/(n_h - 1)\}$, which is itself bounded by $\max_h \{1/((n_h - 1)\}$. If all $n_h > 5$, then the relative bias in v_J is at most 20 percent, which translates into a relative bias in the estimated standard error $\sqrt{v_J}$ of at most 10 percent.

3. The Extended Delete-A-Group Jackknife

In this section, we extend the notion of a DAGJK variance estimator to handle cases where $n_h < R$ for some strata.

Let

w_{hjk} be the weight of element k in PSU j of stratum h

n_h be the number of sampled PSU's in stratum h

H be the number of strata

R be the number of variance groups (the members of each first-stage stratum are distributed into the R replicate groups in as nearly equal a manner as possible)

and

S_{hr} be the set of n_{hr} PSU's in stratum h and group r.

When $n_h < R$, we can define the replicate-r weight of hjk for the *extended delete-a-group* jackknife as

$$w_{hjk(r)}^{(E)} = \begin{cases} w_{hjk} & \text{when } S_{hr} \text{ is empty} \\ w_{hjjk}(1 - [n_h - 1]Z) & \text{when } j \text{ is in } S_{hr}, \text{ and} \\ w_{hjk}(1 + Z) & \text{otherwise,} \end{cases} \tag{2}$$

where $Z^2 = R/[(R-1)n_h(n_h-1)]$. When n_h is larger than or equal to R, $w_{hjk(r)}^{(E)}$ is defined to be the DAGJK replicate weight, $w_{hjk(r)}$; that is, 0 when j is in S_{hr}, and $w_{hjk}n_h/(n_h - n_{hr})$ otherwise. We will see that this leads to a nearly unbiased variance estimation strategy.

When $n_h = R$ in Equation (2), one (and only one) j will be in S_{hr}, $Z = 1/(R-1) = 1/(n_h - 1)$, and the usual DAGJK replicate-weight formula obtains. Observe than when $n_h < R$, the $w_{hjk(r)}^{(E)}$ in Equation (2) is not zero for j in S_{hr}. This is unusual for a jackknife.

To see why the Extended DAGJK works for sufficiently large R, we assume (without loss of generality) that $n_h \le R$ for all h. Let $t = \sum\sum t_{hj}$, and

$$t_{(r)} = \sum_{h \in H^{(r)}} \sum_{j=1}^{n_h} t_{hj} + \sum_{h \in H^r} \left\{ t_{hj^*}(1 - [n_h - 1]Z) + \sum_{j \ne j^*} t_{hj}(1 + Z) \right\}$$

where $H^{(r)}$ is the set of strata with empty S_{hr}, H^r is the set of strata with a PSU in S_{hr}, and j^* is the one PSU in S_{hr}.

Now

$$t_{(r)} - t = \sum_{H^r} \left\{ -t_{hj^*}[n_h - 1]Z + \sum_{j \ne j^*} t_{hj}Z \right\} = -\sum_{H^r} n_h Z \left\{ t_{hj^*} - \sum_{j=1}^{n_h} t_{hj}/n_h \right\}$$

and

$$(t_{(r)} - t)^2 = \sum_{H^r} [R/(R-1)][n_h/(n_h-1)] \left\{ t_{hj^*} - \sum t_{hj}/n_h \right\}^2 + \text{zero-meaned cross terms.}$$

So

$$\begin{aligned} v_J &= [(R-1)/R] \sum^{R} (t_{(r)} - t)^2 \\ &= \sum^{R} \sum_{H_r} [n_h/(n_h-1)] \left\{ t_{hj^*} - \sum t_{hj}/n_h \right\}^2 + \text{zero-meaned cross terms} \\ &= \sum^{H} [n_h/(n_h-1)] \sum^{n_h} (t_{hj} - \sum t_{hi}/n_h)^2 + \text{zero-meaned cross terms,} \end{aligned}$$

because each PSU is in exactly one H^r. This last expression for v_J has the same expectation as the conventional linearization variance estimator assuming with-replacement sampling (or ignoring finite population correction) in the first stage of selection.

4. Discussion

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture has begun using the DAGJK extensively for its surveys. The data from one particular set of multi-phase surveys, the Agriculture Resources and Management Study (ARMS), are shared with the Economics Research Service, a fellow agriculture agency concerned

with economic issues. Consequently, it was imperative to make variance estimation for user-defined complex statistics as simple as possible.

When using a stratified jackknife with a list-based sample, it is a common practice to collapse PSUs within the same stratum into random groups (Wolter 1985, p. 182), although not necessarily the same random groups as for the DAGJK. The ARMS has thousands of PSUs at the national level, so quite a bit of collapsing is warranted before a stratified jackknife can be used. The study also has independent stratification in each of the 48 contiguous states. If the PSUs within each stratum were collapsed into one or two random groups in every state, there would still be hundreds of groups generating a like number of jackknife replicates at the national level. A referee noted that Rust (1985, p. 387) offers a further modification of the stratified jackknife that deletes as few as one random group per stratum, but this does not help in the case of the ARMS.

Sometimes strata are also collapsed for variance estimation purposes. This is a potential source of bias. Moreover, in order to keep the number of jackknife replicates sufficiently large for state-level estimates while still manageable for national estimates, it makes sense to collapse strata across states. The result of these two types of collapsing (PSUs within strata, strata across states) would be a stratified jackknife similar to the DAGJK in appearance.

There is a price to pay for the DAGJK's simplicity, however. NASS, for example, has set R equal to 15 for the ARMS and for its other surveys using the DAGJK. This means that there are only 14 degrees of freedom in univariate t-tests based on DAGJK-derived standard error estimates. Because the replicates are large and nearly equal in size, invoking a t distribution is this context is more defensible than often is the case in survey sampling.

With 14 degrees of freedom, a two-sided 95 percent confidence interval multiplies the estimated standard error by 2.145 rather than the normally-generated 1.96, an increase of slightly less than 10 percent. The slow gains in efficiency to be realized from adding more groups did not seem to justify the increased inconvenience to the agency and the users of its data. For a multivariate analysis, NASS advises users to follow the suggestion of Korn and Graubard (1990) and employ a Bonferroni approach rather than a multivariate F. Here, the loss in efficiency from using a DAGJK is more dearly felt.

5. References

Korn, E.L. and Graubard, B.I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data. The American Statistician, 44, 270–276.

Kott, P.S. (1998). Using the Delete-a-Group Jackknife Variance Estimator in NASS Surveys. RD Research Report No. RD-98-01, USDA, Washington, DC: NASS.

Kott, P.S. and Stukel, D.M. (1997). Can the Jackknife Be Used with a Two-Phase Sample? Survey Methodology, 81–89.

Rust, K. (1985). Variance Estimation for Complex Extimators in Sample Surveys. Journal of Official Statistics, 1, 381–397.

Westat, Inc. (2000). A User's Guide to WesVar, Version 4. Rockville, MD: Westat.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Received October 1999
Revised June 2001