ESTIMATION OF THE EFFECTIVE DEGREES OF FREEDOM IN T-TYPE TESTS FOR COMPLEX DATA

Jiahe Qian, Educational Testing Service
Rosedale Road, MS 02-T, Princeton, NJ 08541

Key Words: Complex sampling, NAEP data, Jackknife, Satterthwaite's formula, Shrinkage estimate

In the analysis of complex survey data, study trends and comparisons usually involve t-type tests and confidence intervals. The National Assessment of Educational Progress (NAEP), which uses a multistage stratified probability sample design, provides examples of these studies. NAEP is designed to assess and report on student academic achievement and educational trends in the United States.

In t-type tests, one basic issue involved is the calculation of the variances of the statistics of interest. Several approaches are often employed in practice such as, interpenetrating subsamples, Jackknife repeated replication (JRR), balanced repeated replication (BRR), bootstrap, and the Taylor series method. In NAEP, the paired JRR is used to estimate variances. Another practical issue in t-type tests is the estimation of the effective degrees of freedom for the variances. Satterthwaite's formula can be used to obtain the effective number of degrees of freedom for the variances estimated by the approaches (Rao & Scott, 1981; Johnson & Rust, 1992) discussed in Section 1.1.

In application, however, some deficiencies in the Satterthwaite's formula would occur. First, it would underestimate the effective degrees of freedom; see Section 1.2. Johnson and Rust (1992) use an adjusted formula that is determined by empirical approach. Second, the measure of the effective degrees of freedom could be unstable; see the discussion in Section 1.3. Examples were found in the statistical tests conducted for the 1996 NAEP long-term trend assessments. To verify the cause of downward bias and instability, a computer simulation was done; see Section 1.4.

The objective of this research is to find approaches that produce stable estimates of the effective degrees of freedom, which are discussed in Sections 2.1-2.2.

This research was partially supported under the National Assessment of Educational Progress (Grant No. R999G50001) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The author thanks Eugene Johnson and Jim Carlson for many helpful comments.

1. THE ESTIMATION OF THE EFFECTIVE DEGREES OF FREEDOM

1.1 The Satterthwaite's Formula

In analysis of independent samples, the number of degrees of freedom for the variance of a linear combination of several estimates is not simply the sum of the numbers of degree of freedom for each estimate. One estimate of the effective number of degrees of freedom is obtained by matching estimates of the first two moments of the variance to those of a chi-square random variable (Satterthwaite, 1941, 1946; Cochran, 1977). The degrees of freedom are the measure of the stability of a variance estimator.

For a linear combination of independent normally distributed $\bar{x} = \sum_{j=1}^{m} \bar{x}_j$ estimates, the effective degree of freedom of the variance of $\bar{x}$ is

$$\hat{df}_S = \left(\sum_{j=1}^{m} S_{x_j}^2\right)^2 / \sum_{j=1}^{m} \left(S_{x_j}^4/(n_j - 1)\right)$$

(Satterthwaite 1941), $S_{x_j}^2 = 1/(n_j - 1)\sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2$, where n_j is the sample size of jth subsample. The effective $\hat{df}_S$ always between the smallest of the $n_j - 1$ and their sum.

For a weighted combination of $\bar{x}_j$,

$$\hat{df}_S = \left(\sum_{j=1}^{m} w_j S_{x_j}^2\right)^2 / \sum_{j=1}^{m} \left(w_j^2 S_{x_j}^4/(n_j - 1)\right).$$

A special case is where the values of n_j are 2. Then $\hat{df}_S = \left(\sum_{j=1}^{m} S_{x_j}^2\right)^2 / \sum_{j=1}^{m} S_{x_j}^4$. Obviously, $1 \le df \le m$. So a smaller $\hat{df}_S$ would cause hypothesis tests and confidence intervals more conservative rather than a traditional one.

The Satterthwaite's formula can be used to calculate the effective number of degrees of freedom for variances estimated by interpenetrating subsamples, Jackknife and BRR approaches.

In the paired Jackknife procedures, two Primary Statistical Units (PSU) are selected from each stratum. In NAEP studies, PSUs usually form 62 pairs. With a paired selection design, one PSU is dropped from stratum 1 at random; then the weights of elements in the other PSU in that stratum are doubled. Based on this set

of weights, statistic $\hat{\theta}_1$, say sample mean, can be calculated to estimate population parameter θ. $\hat{\theta}_1$ is called a pseudo-value. Repeat the process by dropping one PSU from each of the strata in turn, doubling the weights of elements in the other PSU, and computing $\hat{\theta}_2$, $\hat{\theta}_3$,... $\hat{\theta}_m$. Then $V(\hat{\theta})$ is estimated by $v(\hat{\theta}) = \sum_{i=1}^{m} (\hat{\theta}_i - \hat{\theta})^2$. This procedure can readily be extended to designs with more than two sampled PSUs per stratum.

In NAEP samples, the Jackknife variance has the sum of 62 squared variations, which are equivalent to 62 pairs of independent samples in the Satterthwaite's formula. Therefore, the Satterthwaite's formula can be applied to estimate the degrees of freedom for Jackknife variances. Although the Satterthwaite's formula is broadly used in complex data analysis, Satterthwaite estimates have several deficiencies.

1.2 The Adjustment of Downward Bias in Estimation of the Effective Degrees of Freedom

One deficiency in Satterthwaite estimates is the underestimation of the effective degrees of freedom (Johnson & Rust, 1992). It will introduce downward bias in the estimation and cause tests to be too conservative. In NAEP data, the estimated effective degrees of freedom for the NAEP Jackknife variance are sometimes noticeably smaller than the degrees of freedom attributed to the corresponding error estimates from conventional techniques that assume a simple random sampling of students. When comparing the unadjusted $\hat{df}_S$ from the data in Table 2.1 with the average $\hat{df}_S$ in simulation in Table 1.1, the downward bias is clear. The simulation results show that the downward bias could be caused by a violation of the normality assumption. We found that the means and medians of $\hat{df}_S$ when PSU means form a gamma distribution are smaller than those with PSU means normally distributed.

To remedy the downward bias, Johnson and Rust (1992) proposed an adjustment to $\hat{df}_S$

$$\hat{df}_A = \left(3.16 - \frac{2.77}{\sqrt{m}}\right) \frac{\left(\sum_{j=1}^{m} \hat{\sigma}_j^2\right)^2}{\sum_{j=1}^{m} \hat{\sigma}_j^4}.$$

1.3 The Instability of the Estimates of the Effective Degrees of Freedom

Another drawback is the instability in estimation by the Satterthwaite's formula. For example, the instability was found in the report on the 1996 NAEP 8th Grade Mathematics Long-Term Trend Across Assessments. Table 1.2 are the estimated effective degree of freedom for weighted proficiency scores, which are major measurements in NAEP studies.

In Table 1.2 shows that the effective degree of freedom was 14.3 for male students in the 1992 NAEP mathematics long-term trend across assessments, and 43.8 for female students. The largest effective degree of freedom for male students is 34.1 in 1990, which is more than twice that of 14.3 in 1992. The coefficients of the variation are 0.3 and 0.4 for male and female students, respectively. The larger coefficients of variation show instability. In general, the coefficients of variation are greater than or equal to 0.2. Similar problems were found in other NAEP data.

The problems of downward bias and instability may cause errors in statistical results and misleading in decisions, which are unacceptable.

1.4 The Empirical Distribution of the Effective Degrees of Freedom

The problems of underestimation and instability could be caused by the following: First, in empirical survey data, the assumption of normality for the Satterthwaite formula could be violated. Second, the estimate for the effective degrees of freedom is a ratio of two high order moment estimates, so the variance of Satterthwaite's estimates would be very large. Third, the sampling methods within each PSU are usually complex, and therefore, could introduce weighing and design effects.

To verify the causes of the problems, we used a Monte Carlo simulation to approximate the distribution of the estimates of the effective degrees of freedom. Although normality is the assumption for the Satterthwaite's formula, the normal, gamma, and uniform distributions of random variables were employed in the simulation. The simulation was repeated 2000 times. In a typical simulation, the number of random variables, group size, is set at 62, which is the same as the number of Jackknifing replicates in NAEP. The empirical distributions (see Figures 1-3) show that the effective degrees of freedom were distributed close to a normal one.

Table 1.1 lists the means of the degrees of freedom and their standard errors for a typical simulation. When comparing them with the Satterthwaite's estimates from NAEP data, in the first column in Table 2.1, a downward bias can be easily found. Table 1.1 also shows that the coefficients of variation from the simulation are relative small; however, those derived from NAEP mathematics long-term trend across assessments in Table 1.2 are large. These show the evidence of instability of Satterthwaite's estimates. The Figure 4 shows the linear relationship between degrees

of freedom and group sizes.

2. The Improvement in Instability of Satterthwaite Estimates

Several improvements are proposed here to solve the problem of instability.

2.1 A Moderate Number of Degree of Freedom

To remedy the instability, the mean or median estimate of the effective degrees of freedom, $\hat{df}_\mu$, can be used as the effective degrees of freedom in hypothesis tests. We can obtain $\hat{df}_\mu$ by a computer simulation.

2.2 Composite Estimator

To improve the accuracy of estimates for the effective degrees of freedom, a general composite estimator can be expressed as

$$\hat{df}_J = \alpha \hat{df}_\mu + (1 - \alpha)\hat{df}_A,$$

where $\hat{df}_\mu$ is the mean of the simulation distribution and $\hat{df}_A$ is the improved estimate of the effective degrees of freedom (Johnson & Rust, 1991).

2.3 Optimal Shrinkage Estimator

To treat $\hat{df}_A$ as "model-based" estimator, an optimal composite estimator can be defined:

$$\alpha_{CS} \triangleq \begin{cases} A_m/(A_u + A_m) & \text{if } A_m \geq 0 \text{ and } A_u \geq 0, \\ 1 & \text{if } A_u < 0, \\ 0 & \text{otherwise,} \end{cases}$$

with $A_m \triangleq V(\hat{df}_A) + Bias^2(\hat{df}_A)$ and $A_u \triangleq V(\hat{df}_\mu)$. This is a special case of the optimal composite estimator proposed by Cohen and Spencer (1991). In calculation, α_{CS} can be estimated by sample moments for A_m and A_u. And $V(\hat{df}_A)$ would be estimated by the Jackknife approach.

This shrinkage estimator was used to estimate the effective degrees of freedom for the mean scale scores for NAEP assessments. The results of the shrinkage estimates are listed in the second and third columns in Table 2.1. They are stable, and also avoid overadjustments, which sometimes are caused by the adjustments (Johnson & Rust, 1991).

References

Cochran, W. G. (1977). *Sampling Techniques,* 3rd ed. John Wiley & Sons, New York.

Cohen, T. & Spencer, B. (1991). "Shrinkage Weights for Unequal Probability Samples." *Proceedings of the Section on Survey Research Methods*, 625-630.

Johnson, E. & Rust, K. (1992). "Effective Degrees of Freedom for Variance Estimates from a Complex Sample Survey," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Qian, J., & Spencer, B. (1993). "Optimally Weighted Means in Stratified Sampling," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 863-866.

Rao, J.N.K., & Scott, A.J. (1981), "The Analysis of Categorical Data From Complex Sample Surveys: Chi -Squared Tests for Goodness of Fit and Independence in Two-Way Tables," *Journal of the American Statistical Association*, 76, 221-230.

Satterthwaite, F. E. (1941). "Synthesis of Variance," *Psychometrika*, 16, 5, 309-316.

Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components," *Biometrics*, 2, 110-114.

Table 1.1 The Means, Medians and Standard Errors for Three Distributions
(In Simulation: N=2000; Group Size: m=62)

Distribution	Mean	Median	STD
Normal: N(0, 1)	22.10	22.39	3.72
Gamma: G(2, 1)	15.17	14.86	5.47
Uniform: U(0, 1)	33.98	33.97	2.93

Table 1.2 The Estimated Effective DF for the Variance of Mean Proficiency Scores in NAEP Mathematics Long-Term Trend Across Assessments (Age 17)

	1978	1982	1986	1990	1992	1994	1996	CV
Sex								
Male	27.4	28.0	23.0	34.1	14.3	33.0	20.7	0.3
Female	25.9	22.1	19.3	39.5	43.8	18.6	12.1	0.4
Ethnicity								
White	30.0	19.0	17.8	24.5	28.6	25.2	23.3	0.2
Black	30.8	18.4	18.5	10.7	26.8	24.5	29.8	0.3
Hispanic	6.8	7.1	6.4	12.8	9.1	7.0	19.5	0.5
Other	17.0	3.3	6.2	10.6	21.7	6.3	5.0	0.6
Region								
Northeast	16.1	7.2	8.8	28.2	11.3	16.6	12.9	0.4
Southeast	8.0	6.0	14.2	9.0	14.6	12.4	15.5	0.3
Central	6.0	5.1	14.4	13.6	7.8	9.1	26.0	0.6
West	4.4	5.2	7.3	23.5	9.7	6.7	9.7	0.6
Taken Computer Pgm								
Have	7.8	26.2	11.0	28.9	28.6	21.5	21.2	0.4
Have not	31.4	21.6	26.7	33.8	39.8	32.9	38.2	0.2

Table 2.1 The Composite Estimates of DF for the Variance for Mean Proficiency Scores in 1996 NAEP Eighth Grade Mathematics Assessments

	Not adjusted $\hat{df}_s$	Composite estimate for not adjusted $\hat{df}_s$	Adjusted $\hat{df}_s$	Composite estimate for adjusted $\hat{df}_s$
Sex				
Male	11.6	21.0	32.5	23.1
Female	15.6	20.6	43.9	22.7
Ethnicity				
White	9.9	21.1	27.9	23.5
Black	6.5	21.3	18.3	20.6
Hispanic	6.1	21.3	17.1	20.4
Asian	7.8	21.2	22.0	22.0
Region				
Northeast	3.8	21.4	10.7	21.0
Southeast	5.6	21.3	15.8	20.8
Central	9.0	21.1	25.3	23.8
West	8.2	21.2	22.9	22.8

Figure 2. The Distribution of Degree of Freedom
Proficiency Has A Gamma Dist.

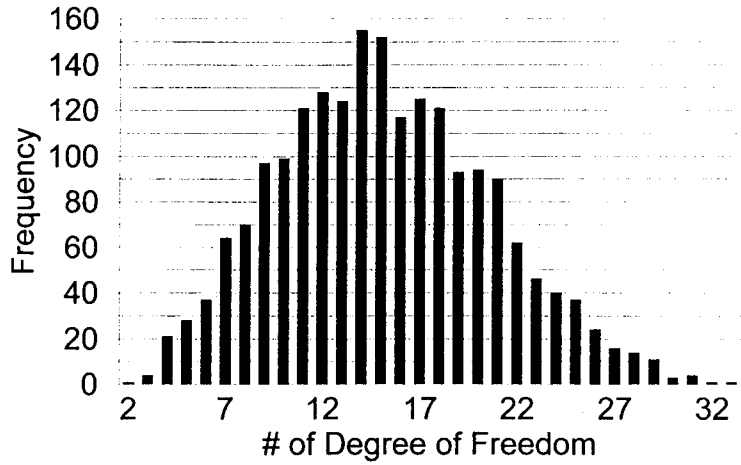

Figure 1. The Distribution of Degree of Freedom
Proficiency is Normally Distributed

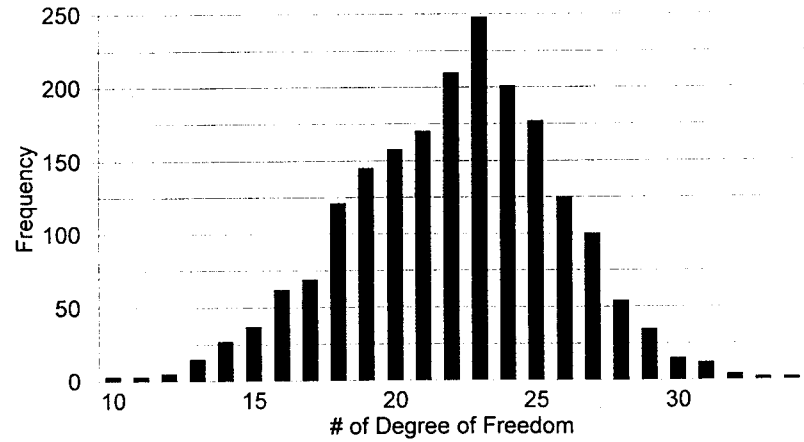

Figure 4. Relationship between DF & Group Sizes
Proficiency is Normally Distributed

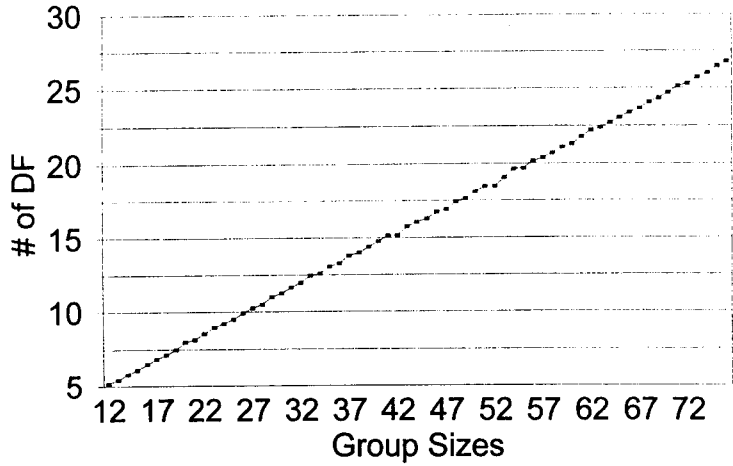

Figure 3. The Distribution of Degree of Freedom
Proficiency is Uniformly Distributed