Journal of Econometrics 3 (1975) 395–404. © North-Holland Publishing Company

ESTIMATION OF VARIANCE AFTER A PRELIMINARY TEST OF HOMOGENEITY AND OPTIMAL LEVELS OF SIGNIFICANCE FOR THE PRE-TEST*

T. TOYODA

Kobe University, Rokkodai-cho, Nadu-ku, Kobe, Japan

T.D. WALLACE

Duke University, Durham, N.C. 27706, U.S.A.

Received May 1974, revised version received April 1975

The question of whether to pool two samples in variance estimation is often decided via a preliminary F test. In this paper we show that the optimal pre-test F value is unity for a one-sided alternative, where the objective function is to minimize average relative risk. The outcome is independent of numbers of degrees of freedom in each sample. Optimal significance levels vary somewhat but are close to $\frac{1}{2}$ for most d.f. and equal to $\frac{1}{2}$ when numerator and denominator d.f. are equal. The results also apply to regression variance estimation across two data regimes.

1. Introduction

Let X_{ij} $(j = 1, 2, \ldots, T_i)$ be a random sample of T_i independent observations drawn from $N(\mu_i, \sigma_i^2)$ with unknown mean μ_i and variance $\sigma_i^2 (i = 1, 2)$. We consider the problem of estimating a parameter σ_1^2 under the condition that the second sample of size T_2 is suspected to have come from the same population as the first. Clearly,

$$s_i^2 = \frac{1}{n_i} \sum_{j=1}^{T_i} (X_{ij} - \bar{X}_i)^2 \tag{1}$$

is an unbiased estimator of σ_i^2 based on $n_i (= T_i - 1)$ degrees of freedom $(i - 1, 2)$. Let us call s_1^2 the *never-pool estimator*. On the other hand, the *always-pool* estimator is defined as

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}. \tag{2}$$

*This work was made possible by a grant from the National Science Foundation, GS30083.

When $\sigma_1^2 = \sigma_2^2$, the always-pool estimator is the appropriate choice of the two and when the variances are different the pooled estimator is biased. However, for at least some values of σ_1^2 and σ_2^2, the pooled estimator will have smaller variance, even though $\sigma_1^2 \neq \sigma_2^2$. In the current problem, however, we assume no a priori knowledge about homogeneity of the variances.

We first conduct a preliminary test for homogeneity of the variances. The statistic for the pre-test is s_2^2/s_1^2, which is distributed according to a central F-distribution with n_2 and n_1 degrees of freedom under the null hypothesis. Assuming that $\sigma_2^2 \geqq \sigma_1^2$ allows us to concentrate on the one-tailed test although the two-tailed test would evoke a similar line of argument. The *sequential* or *sometimes-pool* estimator of σ_1^2 is defined as

$$S^{2*} = \begin{cases} s_1^2, & \text{if } s_2^2/s_1^2 \geqq \lambda, \\ S^2, & \text{if } s_2^2/s_1^2 < \lambda, \end{cases} \tag{3}$$

where λ is an F-value corresponding to an α-level of significance for the preliminary test.

Bancroft (1944) first dealt with this problem; he derived the bias and the variance of S^{2*} and examined the mean square errors of two small sample cases for various values of $\lambda \in [0, \infty]$. Paull (1950) considered a similar problem in the context of the analysis of variance; he suggested a strategy in which the two mean squares are pooled only if their ratio is less than twice the 50% point. Carrillo (1969) investigated the sequential problem, based on a pre-test, of estimating a function of the form $\xi = \sum_{i=1}^{k} w_i \sigma_i^2$, where the w_i's are known weights; he derived biases and mean square errors and compared relative efficiencies for several cases of alternative hypotheses, and suggested some applications to survey sampling. In recent years some advances have been made in choosing some explicit and clear criteria to set optimal levels of significance for preliminary tests in other contexts. For example, Han and Bancroft (1968), for the case of pooling two means when variance is unknown, proposed a procedure to select a significance level which ensures higher relative efficiency than some preassigned value. Sawa and Hiromatsu (1971), Wallace and Ashar (1972), Brook (1972), Bock, Yancey and Judge (1973) and others, for the case of sequential estimation of regression coefficients, have exposed and derived the implications of such criteria as minimax loss, minimax regret and Bayesian minimum expected loss.

The main purpose of this paper is to set optimal significance levels for the preliminary test of variance homogeneity based on a reasonable but tractable criterion. In section 2 the bias and the mean square error of the sequential estimator is presented. Relative efficiency of the estimator and determination of optimal levels of significance for the preliminary test are studied in section 3. Numerical results are discussed in section 4, followed by some concluding remarks.

2. Bias and mean square error

Following Bancroft (1944), the bias of the sequential estimator can be written as

$$Bias\ (S^{2*}) = \frac{n_2}{n_1+n_2}\left[I_{y_0}\left(\frac{n_2}{2}+1, \frac{n_1}{2}\right)-\theta I_{y_0}\left(\frac{n_2}{2}, \frac{n_1}{2}+1\right)\right]\sigma_2^2, \tag{4}$$

where $I_{y_0}(\cdot, \cdot)$ is the incomplete beta function, $y_0 = n_2\lambda\theta/(n_1+n_2\lambda\theta)$, and $0 \leqq \theta = \sigma_1^2/\sigma_2^2 \leqq 1$. Eq. (4) is slightly different from Bancroft's result because it is assumed here that $\sigma_2^2 \geqq \sigma_1^2$ rather than the reverse. The result may be partly checked by noting that when $\lambda = \infty$, i.e., when the two mean squares are always pooled, y_0 becomes 1 and eq. (4) reduces to $n_2(\sigma_2^2-\sigma_1^2)/(n_1+n_2)$. Similarly, when $\lambda = 0$, in which case there is no pooling, y_0 becomes 0 and eq. 4 vanishes.

The mean square error of the estimator S^{2*} is given by

$$MSE(S^{2*}) = E[(S^{2*})^2]-[E(S^{2*})]^2+[Bias\ (S^{2*})]^2. \tag{5}$$

Again, following Bancroft (1944), (5) turns out to be

$$\begin{aligned} MSE(S^{2*}) = \Bigg[&\left\{\frac{2}{n_1}+\frac{2n_2}{n_1+n_2}I_{y_0}\left(\frac{n_2}{2}, \frac{n_1}{2}+1\right)-\frac{n_2(n_1+2)(2n_1+n_2)}{n_1(n_1+n_2)^2}\right. \\ &\left.\times I_{y_0}\left(\frac{n_2}{2}, \frac{n_1}{2}+2\right)\right\}\theta^2+\left\{\frac{2n_1n_2}{(n_1+n_2)^2}I_{y_0}\left(\frac{n_2}{2}+1, \frac{n_1}{2}+1\right)\right. \\ &\left.-\frac{2n_2}{n_1+n_2}I_{y_0}\left(\frac{n_2}{2}+1, \frac{n_1}{2}\right)\right\}\theta+\frac{n_2(n_2+2)}{(n_1+n_2)^2} \\ &\times I_{y_0}\left(\frac{n_2}{2}+2, \frac{n_1}{2}\right)\Bigg]\sigma_2^4, \end{aligned} \tag{6}$$

where y_0 and θ are the same as defined before. Again note that the result is different from Bancroft's because we are assuming the opposite alternative, i.e., $\sigma_2^2 \geqq \sigma_1^2$. In what follows, the mean square error is expressed as a fraction of σ_2^4 so that one unknown parameter is eliminated. If $\lambda \to \infty$, $MSE(S^{2*})/\sigma_2^4$ converges to $[(2n_1+n_2^2)\theta^2-2n_2^2\theta+n_2(n_2+2)]/(n_1+n_2)^2$, which is the mean square error of the always-pool estimator S^2 expressed as a fraction of σ_2^4. If $\lambda \to 0$, $MSE(S^{2*})/\sigma_2^4$ converges to $2\theta^2/n_1$, which is the mean square error of the never-pool estimator s_1^2 expressed as a fraction of σ_2^4.

3. Efficiency and an optimality criterion

From the results given above, $MSE(s_1^2)/\sigma_2^4$ and $MSE(S^2)/\sigma_2^4$ always have two intersections with respect to θ for any values of λ and degrees of freedom, n_1 and

n_2, provided that $n_1n_2-4n_1-2n_2 \neq 0$. The roots are

$$\theta_1 = [n_1n_2-\sqrt{2n_1(n_1n_2+n_2^2+4n_1+2n_2)}]/(n_1n_2-4n_1-2n_2), \quad (7)$$

and

$$\theta_2 = [n_1n_2+\sqrt{2n_1(n_1n_2+n_2^2+4n_1+2n_2)}]/(n_1n_2-4n_1-2n_2). \quad (8)$$

It is easily shown that there are two cases: (i) $0 < \theta_1 < 1, 1 < \theta_2$, and (ii) $0 < \theta_1 < 1, \theta_2 < 0$. In any case there exists one intersection whose root is $\theta_1 \in (0, 1)$.

$MSE(S^{2*})/\sigma_2^4$ moves depending on two parameters, λ and θ, for any given degrees of freedom. As mentioned in section 2, it approaches $MSE(s_1^2)/\sigma_2^4$ and $MSE(S^2)/\sigma_2^4$ accordingly as $\lambda \to 0$ and $\lambda \to \infty$, respectively. Also,

$$\min [MSE(s_1^2)/\sigma_2^4, MSE(S^2)/\sigma_2^4] = \begin{cases} MSE(s_1^2)/\sigma_2^4, & \text{if } 0 \leqq \theta \leqq \theta_1, \\ MSE(S^2)/\sigma_2^4, & \text{if } \theta_1 \leqq \theta \leqq 1. \end{cases} \quad (9)$$

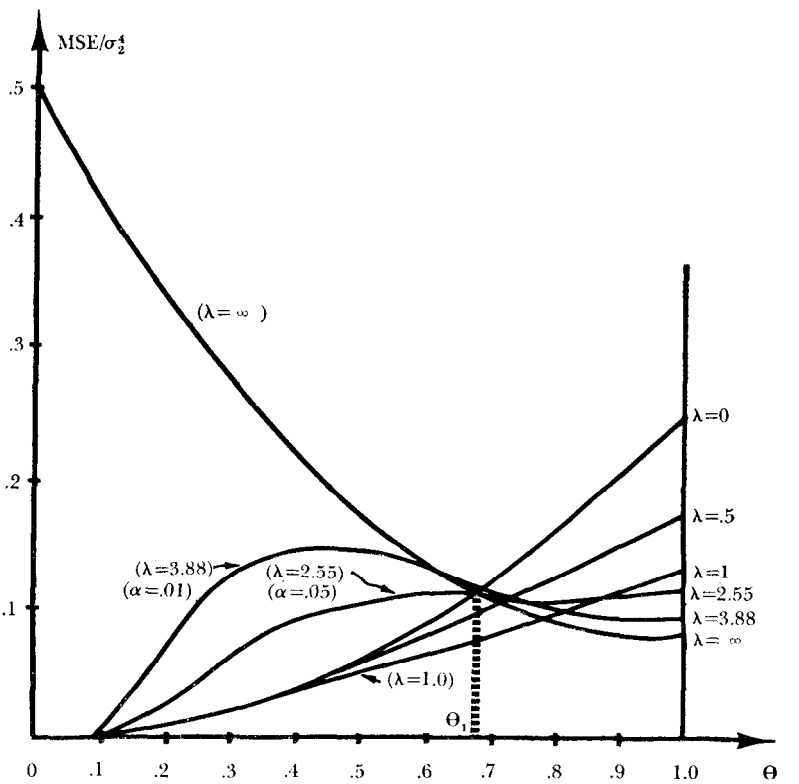

Fig. 1. $MSE/\sigma_2{}^4$ for $n_1 = 8$ and $n_2 = 16$.

As illustrated in fig. 1, there exist some values of $\lambda \in (0, 2)$ such that $MSE(S^{2*}) \leqq MSE(s_1^2)$ over all range of θ, $0 \leqq \theta \leqq 1$. For example, the curve with $\lambda = 1.0$ dominates not only $MSE(s_1^2)/\sigma_2^4$ always but also the curves with $\lambda = 0.05$ and 1.05 almost always which also dominate $MSE(s_1^2)/\sigma_2^4$. Although the curve with $\lambda = 1.0$ does not always dominate $MSE(S^2)/\sigma_2^4$, it does so up to some value of θ around 0.08. For the values of $\lambda \geqq 2.0$, the larger the value of λ, the greater the area represented by the two curves, $\min[MSE(s_1^2)/\sigma_2^4, MSE(S^2)/\sigma_2^4]$ and $MSE(S^{2*})/\sigma_2^4$. These facts indicate that some value of λ around 1.0 would be a reasonable choice for the pre-test. In what follows, we will set an explicit optimality criterion which is reasonable and computationally tractable for choosing λ. In particular, we will analytically show that $\lambda = 1.0$ actually satisfies the optimality criterion not only for the above special case but also in general.

First, define the efficiency of the sequential estimator with respect to the always-pool and the never-pool estimators expressed as a fraction of σ_2^4 as

$$\min [MSE(s_1^2), MSE(S^2)] - MSE(S^{2*}). \tag{10}$$

Fig. 2 exhibits the relative efficiency for the same special case as in fig. 1,

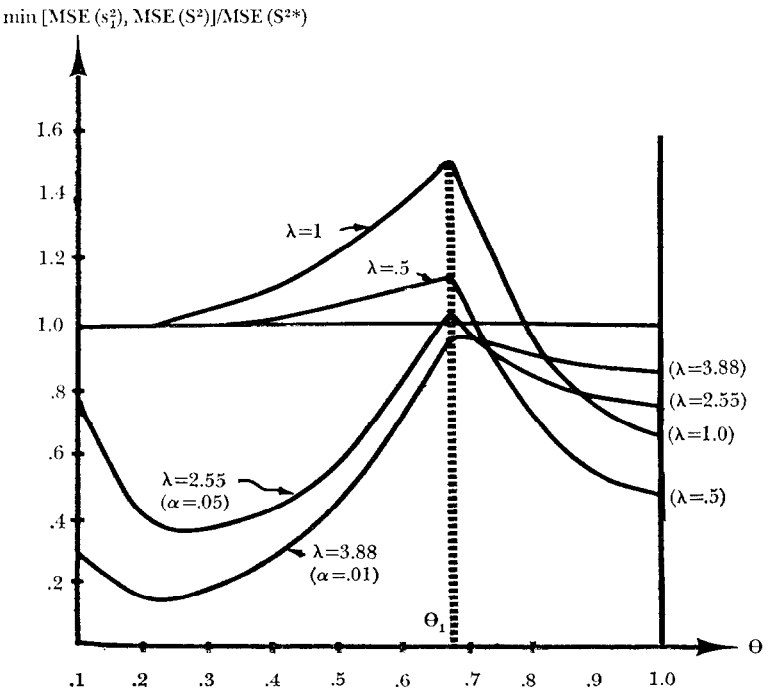

Fig. 2. Relative efficiency of S^{2*} to min $[s_1^2, S^2]$ for $n_1 = 8, n_2 = 16$.

taken as the ratio

$$\min\,[MSE(s_1^2),\ MSE(S^2)]/MSE(S^{2*}),$$

instead of the difference between the numerator and the denominator like (10). It also indicates that $\lambda = 1.0$ gives higher relative efficiency than others over fairly broad range of θ.

Our decision criterion for choosing λ is to maximize the efficiency over the whole range of θ, i.e.,

$$\max_{\lambda} G(\lambda) = \max_{\lambda} \int_0^1 \{\min[MSE(s_1^2),\ MSE(S^2)] - MSE(S^{2*})\}\,\mathrm{d}\theta. \quad (11)$$

Note that this is equivalent to maximizing the average efficiency provided that the prior distribution of θ is diffuse,[1] $G(\lambda)$ can be rewritten as

$$\begin{aligned}
G(\lambda) &= \int_0^{\theta_1} [MSE(s_1^2) - MSE(S^{2*})]\,\mathrm{d}\theta \\
&\quad + \int_{\theta_1}^1 [MSE(S^2) - MSE(S^{2*})]\,\mathrm{d}\theta \\
&= \int_{\theta_1}^1 \left[\left\{\frac{2n_1+n_2^2}{(n_1+n_2)^2} - \frac{2}{n_1}\right\}\theta^2 - \frac{2n_2^2}{(n_1+n_2)^2}\,\theta + \frac{n_2(n_2+2)}{(n_1+n_2)^2}\right]\mathrm{d}\theta \\
&\quad + \int_0^1 \left[\left\{\frac{n_2(n_1+2)(2n_1+n)^2}{n_1(n_1+n_2)^2} I_{y_0}\left(\frac{n_2}{2}, \frac{n_1}{2}+2\right)\right.\right. \\
&\quad \left. - \frac{2n_2}{n_1+n_2} I_{y_0}\left(\frac{n_2}{2}, \frac{n_1}{2}+1\right)\right\}\theta^2 \\
&\quad - \left\{\frac{2n_1n_2}{(n_1+n_2)^2} I_{y_0}\left(\frac{n_2}{2}+1, \frac{n_1}{2}+1\right) - \frac{2n_2}{n_1+n_2} I_{y_0}\left(\frac{n_2}{2}+1, \frac{n_1}{2}\right)\right\}\theta \\
&\quad \left. - \frac{n_2(n_2+2)}{(n_1+n_2)^2} I_{y_0}\left(\frac{n_2}{2}+2, \frac{n_1}{2}\right)\right]\mathrm{d}\theta. \qquad (12)
\end{aligned}$$

[1] If we have a priori information about the distribution of θ, we can of course obtain different optimal significance levels from the ones stated below. For a related discussion in pooling means using prior information, see Han and Bancroft (1968). Also note that $\{\lambda^* \mid \max G(\lambda^*)\} = \{\lambda^* \mid \min H(\lambda^*)\}$, where $H(\lambda)$ is the area under the curve represented by $MSE(S^{2*})/\sigma_2{}^4$.

Using integrations by parts and making necessary substitutions, we obtain the contracted form of $G(\lambda)$, which is

$$G(\lambda)=\frac{1}{3}\left\{\frac{2n_1+n_2^2}{(n_1+n_2)^2}-\frac{2}{n_1}\right\}(1-\theta_1^3)-\frac{n_2^2}{(n_1+n_2)^2}(1-\theta_1^2)$$

$$+\frac{n_2(n_2+2)}{(n_1+n_2)^2}(1-\theta_1)+\frac{n_2(n_1+2)(2n_1+n_2)}{3n_1(n_1+n_2)^2}\left]_{\frac{n_2\lambda}{n_1+n_2\lambda}}\left(\frac{n_2}{2},\frac{n_1}{2}+2\right)\right.$$

$$-\frac{2n_2}{3(n_1+n_2)}\left]_{\frac{n_2\lambda}{n_1+n_2\lambda}}\left(\frac{n_2}{2},\frac{n_1}{2}+1\right)\right.$$

$$-\frac{n_1n_2}{(n_1+n_2)^2}\left]_{\frac{n_2\lambda}{n_1+n_2\lambda}}\left(\frac{n_2}{2}+1,\frac{n_1}{2}+1\right)\right.$$

$$+\frac{n_2}{n_1+n_2}\left]_{\frac{n_2\lambda}{n_1+n_2\lambda}}\left(\frac{n_2}{2}+1,\frac{n_1}{2}\right)-\frac{n_2(n_2+2)}{(n_1+n_2)^2}\right]_{\frac{n_2\lambda}{n_1+n_2\lambda}}$$

$$\times\left(\frac{n_2}{2}+2,\frac{n_1}{2}\right)+\frac{n_1(n_2+2)(n_2+4)(3\lambda n_1+3\lambda^2n_2-2n_1-n_2)}{3\lambda^3n_2(n_1-2)(n_1+n_2)^2}$$

$$\times\left]_{\frac{n_2\lambda}{n_1+n_2\lambda}}\left(\frac{n_2}{2}+3,\frac{n_1}{2}-1\right)+\frac{n_1^2(n_2+2)(n_2+4)(2-3\lambda)}{3\lambda^3n_2(n_1+n_2)(n_1-4)(n_1-2)}\right.$$

$$\times\left]_{\frac{n_2\lambda}{n_1+n_2\lambda}}\left(\frac{n_2}{2}+3,\frac{n_1}{2}-2\right).\right. \tag{12'}$$

The reader may refer to an appendix for the algebra resulting in the above expression for $G(\lambda)$.

After some calculations we find

$$\left.\frac{\mathrm{d}G(\lambda)}{\mathrm{d}\lambda}\right|_{\lambda=1}=0, \tag{13}$$

which shows that the necessary condition for the maximum is attained when $\lambda = 1$. To assure the sufficiency and the uniqueness of the maximum, we have conducted iterative computations of $G(\lambda)$ for various degrees of freedom, beginning from $\lambda = 1.00$ and giving increments by ± 0.01 for each case.

4. Type one errors of the optimal pre-test and average maximum efficiencies

In the preceding section it was shown that maximizing relative average

efficiency of a pre-test estimator of a variance leads to a critical value of unity for the preliminary test; i.e., $\lambda = 1$. Table 1 gives the corresponding values for type one error (say α^*) and values of the maximum average efficiency, $G(1)$, for selected n_1 and n_2. For instance, if $n_1 = 24$ and $n_2 = 8$, $\lambda = 1$ implies a type one error for the pre-test of $\alpha^* = 0.539$. The corresponding maximum average efficiency is 1.01×10^{-3}.

The table shows stable values for α^* over substantial variation in n_1 and n_2, the optimal significance level is always $\frac{1}{2}$. Other α^* values vary from about 0.4 to 0.6. For $n_1 > n_2$, the optimal levels are greater than $\frac{1}{2}$ and the reverse is true for $n_1 < n_2$.

No doubt there exist continuous weight function estimators (of the test statistic) which dominate the pre-test estimator for any choice of critical value for the pre-test. However, for those who continue to prefer pre-testing procedures for whatever reasons, it is interesting that there exists a unique value of α that minimizes relative risk averaged over the range of the nuisance parameter, regardless of degrees of freedom. Since the results given here are in a sense the best one can do in pre-test estimation for variances, they at least provide a benchmark for evaluating the benefits of Stein type estimation in such cases.

The results given here apply directly to estimating regression variances across two data sets. For regression variances, the n_1, n_2 in table 1 would refer to degrees of freedom in the two regressions.

Table 1
Values of α^* and $G(1)$.

n_2 \ n_1		8	16	24	60	120
2	α^*	0.590	0.610	0.617	0.626	0.629
	$G(1)$	6.86×10^{-3}	1.96×10^{-3}	9.15×10^{-4}	1.55×10^{-4}	3.96×10^{-5}
4	α^*	0.539	0.564	0.573	0.585	0.590
	$G(1)$	6.79×10^{-3}	2.17×10^{-3}	1.06×10^{-3}	1.90×10^{-4}	4.95×10^{-5}
8	α^*	0.500	0.527	0.539	0.554	0.560
	$G(1)$	4.95×10^{-3}	1.91×10^{-3}	1.01×10^{-3}	2.04×10^{-4}	5.56×10^{-5}
16	α^*	0.473	0.500	0.512	0.531	0.538
	$G(1)$	1.71×10^{-3}	1.06×10^{-3}	6.69×10^{-4}	1.74×10^{-4}	5.25×10^{-5}
24	α^*	0.461	0.488	0.500	0.520	0.528
	$G(1)$	-3.49×10^{-4}	3.45×10^{-4}	3.31×10^{-4}	1.27×10^{-4}	4.35×10^{-5}
60	α^*	0.446	0.469	0.481	0.500	0.510
	$G(1)$	-4.18×10^{-3}	-1.33×10^{-3}	-5.99×10^{-4}	-6.14×10^{-5}	-2.90×10^{-6}
120	α^*	0.440	0.462	0.472	0.490	0.500
	$G(1)$	-5.98×10^{-3}	-2.28×10^{-3}	-1.21^{-3}	-2.39×10^{-4}	-5.86×10^{-5}

Appendix

Finding $G(\lambda)$ requires evaluation of an integral of the form

$$A_i = \int_0^1 \theta^i I_{y_0}(\alpha, \beta)\, d\theta, \qquad i = 0, 1, 2, \tag{A.1}$$

where $I_{y_0}(\alpha, \beta)$ is the incomplete beta function and $y_0 = n_2\lambda\theta/(n_1+n_2\lambda\theta)$.

Integrating by parts,

$$A_i = \frac{1}{i+1}\left[I_{x_0}(\alpha, \beta) - \int_0^1 \theta^{i+1} I'_{y_0}(\alpha, \beta)\, d\theta\right], \tag{A.2}$$

where $x_0 = n_2\lambda/(n_1+n_2\lambda)$.

But

$$I'_{y_0}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y_0^{\alpha-1}(1-y_0)^{\beta-1}\frac{dy_0}{d\theta},$$

so

$$A_i = \frac{1}{i+1}\left[I_{x_0}(\alpha, \beta) - \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{n_1}{n_2\lambda}\right)^{i+1}\int_0^{x_0} y_0^{\alpha+1}(1-y_0)^{\beta-i-2}dy_0\right]. \tag{A.3}$$

Hence,

$$A_i = \frac{1}{i+1}\left[I_{x_0}(\alpha, \beta) - \left(\frac{n_1}{n_2\lambda}\right)^{i+1}\frac{\Gamma(\alpha+i+1)\Gamma(\beta-i-1)}{\Gamma(\alpha)\Gamma(\beta)} \times I_{x_0}(\alpha+i+1, \beta-i-1)\right], \tag{A.4}$$

for $i = 0, 1, 2$. Substituting this result into eq. (12) in the text and collecting terms yields the equation for $G(\lambda)$.

References

Bancroft, T.A., 1944, On biases in estimation due to the use of preliminary tests of significance, Annals of Mathematical Statistics 15, 190–204.

Bock, M.E., T.A. Yancey and G.G. Judge, 1973, The statistical consequences of preliminary test estimators in regression, Journal of the American Statistical Association 68, 109–116.

Brook, R.J., 1972, On the use of a minimax regret function to set significance points in prior tests of estimation, unpublished Ph.D. thesis (North Carolina State University, Raleigh).

Carrillo, A.F., 1969, Estimation of variance after preliminary tests of significance, unpublished Ph.D. thesis (Iowa State University, Ames).

Han, C. and T. A. Bancroft, 1968, On pooling means when variance is unknown, Journal of the American Statistical Association 63, 1333–1342.

Paull, A.E., 1950, On a preliminary test for pooling mean squares in the analysis of variance, Annals of Mathematical Statistics 21, 539–556.
Sawa, T. and T. Hiromatsu, 1971, Minimax regret significance points for a preliminary test in regression analysis, Technical Report 39, Institute for Mathematical Studies in the Social Sciences (Stanford University, Stanford, Calif.).
Wallace, T.D. and V.G. Asher, 1972, Sequential methods in model construction, The Review of Economics and Statistics 54, 172–178.