ON A PRELIMINARY TEST FOR POOLING MEAN SQUARES IN THE ANALYSIS OF VARIANCE[1]

By A. E. Paull

Grain Research Laboratory, Winnipeg

Summary. The paper describes the consequences of performing a preliminary F-test in the analysis of variance. The use of the 5% or 25% significance level for the preliminary test results in disturbances that are frequently large enough to lead to incorrect inferences in the final test. A more stable procedure is recommended for performing the preliminary test in which the two mean squares are pooled only if their ratio is less than twice the 50% point.

I. Introduction

The problem discussed in this paper is one of a large class involving preliminary tests of significance. Studies of this type have recently been made by Bancroft [1] and Mosteller [2]. Bancroft dealt with a preliminary test for homogeneity of two variances, and a test of a regression coefficient. Mosteller dealt with the problem of pooling means from two normal populations having the same known variance. The present problem is an extension of Bancroft's work from investigations of the bias and variance of an estimate of variance, to investigations of the consequences of using that estimate in performing a further test of significance.

The problem arises frequently in the analysis of variance. As a simple example, consider an experiment carried out to test the hypothesis that different laboratories in a district all determine the protein content of wheat without systematic differences between laboratories. Three laboratories are selected at random and each is requested to analyze ten samples of the same wheat, five on each of two days. The analysis of variance would be set up in one of two ways:

MODEL I

Source of variation	*d f*	*M S*
Between laboratories	2	v_3
Between days within labs.	3	v_2
Within days	24	v_1

MODEL II

Source of variation	*d f*	*M S*
Between laboratories	2	v_3
Within laboratories	27	$\frac{3v_2 + 24v_1}{27}$

The soundest procedure is to follow Model I in which the F-ratio, v_3/v_2, provides a valid though not very powerful test of the null hypothesis. But the investigator often doubts that this is the most effective form of analysis. His past experience may have shown that measurements of this kind seldom exhibit day-to-day variations appreciably greater than their within-day variations. If he is willing to accept this credible assumption, he adopts Model II because

[1] Based on a doctoral dissertation submitted to the Faculty of North Carolina State College of the University of North Carolina at Raleigh, N. C., in June, 1948. Published as Paper No. 107 of the Grain Research Laboratory, Board of Grain Commissioners, Winnipeg.

this increases the degrees of freedom from 2 and 3 to 2 and 27. These two models may conveniently be called the "never pool" and the "always pool" procedures.

The investigator often prefers what may be called a "sometimes pool" procedure. He starts with Model I and examines the null hypothesis that the variation between days is no greater than the variation within days by testing the F-ratio v_2/v_1. For this test, he selects a probability level P_1 that may be the 5% or some higher level. If the hypothesis of this preliminary test is not rejected, his judgement has been substantiated and he adopts Model II and pcols the two mean squares. If the hypothesis is rejected, he retains Model I since he concludes that v_2 alone is the only valid estimate of error.

The following notation is introduced:

Degrees of freedom	*Mean square*	*Expected value of mean square*
n_3	v_3	σ_3^2
n_2	v_2	σ_2^2
n_1	v_1	σ_1^2

where $\sigma_1^2 \leq \sigma_2^2 \leq \sigma_3^2$.

The mean squares v_1, v_2, and v_3 are assumed to be distributed as central chi-squares. This assumption is justified if the treatments (laboratories in the example) are selected at random from a population of treatments. But if, as is more frequently the case, the experimenter is interested only in specified treatments, the non-central chi-square model is the appropriate one. However, if the two cases are sufficiently parallel, as seems probable, conclusions drawn from the central model may be expected to apply to the non-central model.

Let $\theta_{21} = \sigma_2^2/\sigma_1^2$ and $\theta_{32} = \sigma_3^2/\sigma_2^2$, and let $F(\nu_1, \nu_2, P)$ denote the value exceeded by F for ν_1 and ν_2 degrees of freedom with probability P. The rule of procedure for the "sometimes pool" test may be restated as follows:

Reject the main hypothesis that $\sigma_3^2 = \sigma_2^2 (\theta_{32} = 1)$ if

$$v_2/v_1 \geq F_1(n_2, n_1; P_1) \quad \text{and} \quad v_3/v_2 \geq F_2(n_3, n_2; P_2)$$

or if

$$v_2/v_1 < F_1(n_2, n_1; P_1) \quad \text{and} \quad (n_2 + n_1)v_3/(n_2v_2 + n_1v_1) \geq F_3(n_3, n_2 + n_1; P_3).$$

The "never pool" procedure in which P_2 is used, and the "always pool" procedure in which P_3 is used, may be considered as special cases of the "sometimes pool" procedure in which P_1 takes on its extreme values, 1 and 0 respectively. In practice, the probability levels P_2 and P_3 are usually the same; in the present study they are allowed to be different in case this greater flexibility should prove desirable. The objects of the investigation are: (a) to examine the Type I error under the above rule of procedure, i.e., to determine the frequency of rejecting the null hypothesis when it is true; and (b) to examine the behaviour of the power with particular reference to comparisons with the power of the "never pool" procedure.

The remainder of this paper is divided into four sections: Part II contains a

general discussion of the results, conclusions and recommendations; and Part III illustrates the general conclusions with numerical examples. The derivation of distributions, proofs by elementary arguments of general qualitative results, and derivations of closed form expressions for $n_3 = 2$, are given in Part IV.

II. General Discussion of Results, Conclusions and Recommendations

2.1. Criterion employed. In this part the principal results and recommendations are discussed for the reader who is not interested in the mathematical details. To give results in a simple form is not easy, because of the many variables —the P's, the θ's, and the n's—that enter into the problem. It may be helpful to consider what is wrong with the "always pool" test, and then to state the properties which the preliminary test must have if it is to be regarded as useful and successful.

If the "always pool" procedure is employed when in fact σ_2^2 is greater than σ_1^2, i.e. $\theta_{21} > 1$, the denominator in the final F test tends to be too small. Thus the final F test gives too many significant results when its null hypothesis is true and if θ_{21} is great enough, there is no bound to this hidden distortion of the significance level. A test which the research worker thinks is being made at the 5% level might actually be at, say, the 47% level.

The preliminary test represents an attempt to avoid this alarming disturbance, since if θ_{21} is very large the test is expected to warn against pooling. Such a procedure, however, can not be expected to remove this disturbance completely, and it does not do so, but to be successful it should keep the true or effective significance level of the final F test close to the nominal level at which the research worker thinks he is working.

A second requirement is that the preliminary test should increase the power in the final F test relative to the power of the "never pool" test. When the powers of the "sometimes pool" and "never pool" tests are compared, it is important to make the comparison *at the same significance level.* Suppose the preliminary test shifts the significance level of the final F test from the 5% to the 6% level—a disturbance that for some uses would not be regarded as serious. In this event the "sometimes pool" test (at the 6% level) would tend to be more powerful than the "never pool" test at the 5% level, because an increase in significance level generally results in an increase in power. But unless the "sometimes pool" test has more power than a "never pool" test made also at the 6% level, it has no real advantage over the "never pool" procedure.

2.2. Effect of preliminary tests made at the 5% level. Probably the most common procedure in practice is to perform the preliminary test at the 5% level (i.e. $P_1 = .05$) and, whether pooling is prescribed or not, to conduct the final F test also at the 5% level, (i.e. $P_2 = P_3 = .05$). Such a procedure, except when θ_{21} is near one and the null hypothesis is true, results in the null hypothesis being rejected more frequently than if pooling is never resorted to.

When the ratio θ_{21} is equal to one, so that routine pooling would be valid, the

preliminary test is effective. The true significance level of the final F test is *decreased* slightly, but is always confined between the 5% and the 4.75% levels. Further, the power is always greater than that of the "never pool" test made at the same significance level.

As θ_{21} increases from 1, the true significance level of the final F test increases to a maximum and then slowly decreases to 5%. Unfortunately the maximum need not be near to 5%: in the example presented later it is about 15%, and for a broad range of values of θ_{21} the true significance level is higher than 10%. Comparison with the power of the "never pool" test is also unfavorable to the "sometimes pool" test. For values of θ_{21} near 1, the "sometimes pool" test has the higher power, but as θ_{21} becomes larger the advantage passes to the "never pool" test.

When θ_{21} is very large there is, as would be expected, little disturbance. The preliminary test seldom prescribes pooling, so that the properties of the "sometimes pool" test are very similar to those of the "never pool" test, although the "never pool" procedure yields the slightly higher power.

The main objection to the use of the "sometimes pool" test is associated with the intermediate values of θ_{21}. If over a series of experiments θ_{21} has a moderate value greater than one, the "sometimes pool" test at the 5% levels yields more apparently significant results than are anticipated, and is also less powerful than a corresponding "never pool" test. The magnitude of these undesirable properties can be reduced somewhat by increasing the significance level of the preliminary test.

2.3. Effect of preliminary tests made at the 25% level. Use of the 25% instead of the 5% significance level for the preliminary test reduces, in general, the probability of rejecting the hypothesis. This reduction, at intermediate values of θ_{21}, results in a reduction of the extreme disturbances. When the ratio θ_{21} is equal to one, however, the effects are not as favourable. If the hypothesis is true, still fewer apparently significant results occur. A final test being carried out at the 5% level can now have an effective significance level close to 3.75%. If the hypothesis is false, the test is still more powerful than a corresponding "never pool" test but the gain is not as great as when a preliminary test at the 5% level is employed. Since most experimenters desire a reasonable amount of protection against an error in judgement of the true value of θ_{21}, the reduction in disturbances for intermediate values of θ_{21}, resulting from the use of the 25% rather than the 5% level, would be judged to outweigh the disadvantages of the compensating factors.

2.4. Effect of further increases in significance level. Increasing P_1, the significance level of the preliminary test, decreases the probability of rejecting the hypothesis only to the point where a critical value $\bar{P}_1$ is reached. Increasing P_1 beyond this value results in an increase in the probability of rejection. The properties of a "sometimes pool" test in which P_1 is less than $\bar{P}_1$ differ, in general, from those of a test in which P_1 is greater than $\bar{P}_1$.

Tests of the former type, which are referred to here as Class A tests, are the tests commonly encountered in practice. Considering, for example, a test in which $P_2 = P_3 = .05$ and $n_1 = 20$, $n_2 = 4$, $n_3 = 2$, we find the critical value $\bar{P}_1$ to be .77, a figure much larger than the values .05 or .25 customarily chosen for P_1. The major portion of the present discussion deals with Class A tests. Tests in which P_1 is greater than $\bar{P}_1$ are referred to as Class B tests and discussion of their properties is relegated to a later section. An expression for evaluating $\bar{P}_1$ is given in Subsection 4. 3.

2.5. Effect of P_2, P_3. The probability levels (P_2, P_3) used for the final test determine the properties of the "sometimes pool" test for extreme values of θ_{21}. When θ_{21} is equal to one, the effective significance level is less than the nominal value P_3, but is not less than $(1 - P_1)P_3$. The power of such a test is greater than the power of a corresponding "never pool" test, but less than the power of a test in which one always pools and uses the P_3 level. For very large values of θ_{21} the behavior of the "sometimes pool" test approaches, in all respects, the behaviour of a "never pool" test at the P_2 level.

2.6. Effect of n_2, n_1. The degrees of freedom n_2 and n_1, associated with the mean squares that are sometimes pooled, clearly affect the magnitude of the disturbance. Because analytic investigation becomes complex, the following remarks are based on conjectures arising out of examination of a number of numerical examples.

A large value of n_2 is desirable in two respects. As n_2 becomes larger the preliminary test becomes more powerful and pooling is prescribed less often. In addition, when pooling is prescribed the pooled mean square is further weighted in favour of the valid error σ_2^2. Both factors are contributing towards a decrease in bias of the error mean square with a consequent reduction in the disturbance introduced into the final test.

The effect of n_1 is not as simple. As n_1 becomes larger the preliminary test again becomes more powerful and pooling is prescribed less often. But when pooling is prescribed, the pooled mean square in this case is further weighted in favour of σ_1^2, which is smaller than the valid error σ_2^2. The effect on the final test, which is due to a combination of these two factors, clearly depends on the value of θ_{21}. For intermediate values of θ_{21} the latter factor is the predominant one, and the disturbance of the effective significance level is increased as n_1 is increased.

2.7. Class B Test. A Class B test is one in which the probability level (P_1) of the preliminary test is greater than a critical value $\bar{P}_1$. Pooling is prescribed only when the mean square v_1 is relatively large, with the result that the error mean square tends to be too large. Accordingly, a Class B "sometimes pool" test rejects the hypothesis less frequently than a "never pool" test at the P_2 level.

The effective significance level of a Class B test is less than P_2 for all values of θ_{21}. It has its lowest value when θ_{21} is equal to one, and approaches P_2 as θ_{21}

becomes very large. Because pooling is prescribed infrequently, little power is gained by the use of a Class B test rather than a "never pool" test.

2.8. Recommendations. The principal conclusions discussed in the preceding subsections may be summarized as follows: A preliminary test carried out at a significance level as low as 5% affords little protection against errors in judgement. If σ_1^2 is equal to $\sigma_2^2(\theta_{21} = 1)$ the reduction in errors of inference is appreciable; but if, in fact, σ_1^2 is less than $\sigma_2^2(\theta_{21} > 1)$, a greater number of incorrect inferences are made than if a preliminary test is not employed at all. The use of the 25% significance level for the preliminary test introduces the same disturbances but to a lesser extent. Extreme increases in the effective significance level at possible values of θ_{21} are reduced and losses in power at these values are not as serious. The 25% level provides a reasonable amount of protection against an error in judgement regarding the true value of θ_{21}. However, when n_2 is large relative to n_1, a smaller significance level could be employed without introducing any serious disturbances at the intermediate values of θ_{21}, and with a resulting gain in power at values of θ_{21} near one.

The following method of performing a preliminary test is recommended as one which tends to stabilize the disturbances at intermediate values of θ_{21} while still taking advantage of a considerable portion of the possible gain in power at values of θ_{21} near one. The procedure consists of pooling the two mean squares v_2 and v_1 only if their ratio is less than $2\ F_{50}$, where F_{50} is the 50 per cent point of the F-distribution for n_2 and n_1 degrees of freedom. The use of the multiple 2 is arbitrary and a smaller value may be used if the experimenter desires additional control over extreme disturbances.

This procedure has the advantage of admitting less disturbance over a larger range of values of n_2 and n_1. The customary method prescribes pooling if the null hypothesis ($\theta_{21} = 1$) of the preliminary test is not rejected at some preassigned probability level P_1. If enough observations are available to provide reliable values for v_2 and v_1, pooling is prescribed only if σ_2^2 and σ_1^2 are essentially the same. However, if small numbers of degrees of freedom are involved, the preliminary test is too weak to reject the hypothesis even if σ_1^2 is appreciably less than σ_2^2, and pooling will be prescribed too frequently. On the other hand, the use of the recommended procedure has the effect of prescribing pooling only when it can be said, with confidence exceeding 50%, that the true value of θ_{21} is less than some chosen value such as 2.

This can be demonstrated simply by considering a series of experiments in which preliminary tests are performed. When $v_2/v_1 < 2F_{50}$, we make the statement

$$\theta_{21} < 2, \tag{1}$$

and when $v_2/v_1 \geq 2F_{50}$, we make the statement

$$\theta_{21} \geq 2. \tag{2}$$

We have

$$Pr\left\{\frac{v_2}{v_1} \cdot \frac{1}{\theta_{21}} \geq F_{50}\right\} = .50,$$

or

$$Pr\left\{\frac{v_2}{v_1} \geq F_{50}\theta_{21}\right\} = .50.$$

If statement (1) is true,

$$Pr\left\{\frac{v_2}{v_1} < 2F_{50}\right\} \geq .50;$$

and if statement (2) is true,

$$Pr\left\{\frac{v_2}{v_1} \geq 2F_{50}\right\} \geq .50.$$

Thus, no matter what the true value of θ_{21}, the statements are true more than 50% of the time.

Fifty per cent points of the F-distribution have been tabulated by Merrington and Thompson [3].

A simpler rule, and one which is nearly equivalent when the degrees of freedom involved are each greater than 6, is to pool if the ratio of the mean squares is less than 2, without any reference to the F-table. For smaller numbers of degrees of freedom, however, this simpler rule does not embody the advantages of the $2F_{50}$ rule, unless of course, n_1 and n_2 are equal.

III. Numerical Illustrations

3.1. Effect of P_1 illustrated. An example of the influence of P_1 on the effective significance level or Type I error of a "sometimes pool" test is illustrated in Figure 1. When $P_1 = 0$, the Type I error has its maximum value equivalent to the Type I error of an "always pool" test at the P_3 level. As P_1 increases from zero, the Type I error decreases until at $P_1 = \bar{P}_1$(.77 in this case) it reaches its minimum value at a level less than P_2. As P_1 increases from $\bar{P}_1$, the Type I error increases until, at $P_1 = 1$, the Type I error is equal to P_2.

The influence of P_1 on the power of a "sometimes pool" test is illustrated in Figure 2. The gain in power, as a function of θ_{21} is presented for three Class A tests. Since comparisons of power are made over tests having different Type I errors, the gain is expressed as the proportion actually attained of the total gain in power that is possible if the true value of θ_{21} is actually known. When $P_1 = \bar{P}_1 = .77$, the curve is observed to decrease monotonically to zero. However, for lower values of P_1, the preliminary test prescribes pooling more often, and more power is gained when θ_{21} is near one but less power is gained or power is actually lost when θ_{21} is large.

The power gained or lost at various values of θ_{21} is illustrated in Table I. The probability of rejecting the hypothesis for the "sometimes pool" test is

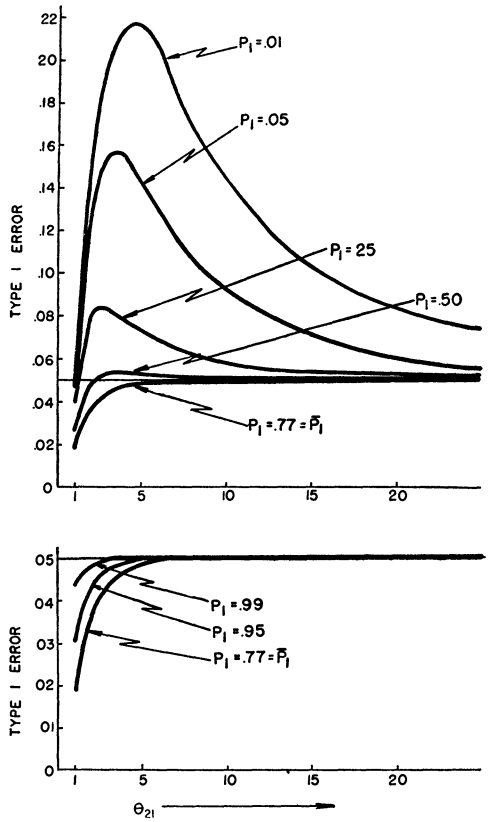

FIG. 1. Effect of Varying P_1. $n_1 = 20$, $n_2 = 4$, $n_3 = 2$ and $P_2 = P_3 = .05$. (a) Upper diagram: Class A Tests. (b) Lower diagram: Class B Tests.

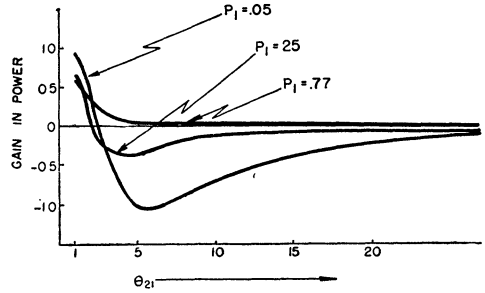

FIG. 2. Proportion of Possible Gain in Power Actually Attained. $n_1 = 20$, $n_2 = 4$, $n_3 = 2$, $P_2 = P_3 = .05$.

tabulated opposite "s.p.", and for the "never pool" test *having the same Type I error* opposite "n.p.".

The last line of the table approaches the probabilities for a "never pool" test having a Type I error of 5%. Except for values very near $(\theta_{21}, \theta_{32}) = (1, 1)$, the probability of rejecting the null hypothesis, using a "sometimes pool" test, is greater than if a "never pool" test, *at the* 5% level is used. In this sense, the

TABLE I

Comparison of Power of a "Sometimes Pool" (s.p.) Test and Corresponding "Never Pool" (n.p.) Tests

$n_1 = 20, n_2 = 4, n_3 = 2; P_1 = P_2 = P_3 = .05$

Value of θ_{21}	Test	Type I Error $\theta_{32} = 1$	Value of θ_{32}							
			1.8	2.8	4.3	7.1	12.5	25	50	250
1.0	s.p.	.048	.164	.299	.443	.599	.739	.855	.922	.984
	n.p.	.048	.112	.192	.297	.441	.604	.765	.870	.972
1.2	s.p.	.067	.200	.338	.476	.621	.751	.860	.925	.984
	n.p.	.067	.149	.245	.361	.508	.662	.805	.895	.978
1.6	s.p.	.102	.248	.379	.503	.632	.750	.855	.921	.983
	n.p.	.102	.210	.323	.447	.592	.730	.849	.920	.983
2.0	s.p.	.127	.271	.390	.500	.619	.736	.845	.915	.981
	n.p.	.127	.250	.370	.497	.636	.764	.870	.932	.986
2.5	s.p.	.146	.278	.382	.482	.596	.715	.831	.907	.975
	n.p.	.146	.278	.402	.528	.664	.784	.882	.938	.987
4.5	s.p.	.148	.233	.309	.399	.520	.657	.796	.887	.976
	n.p.	.148	.280	.405	.531	.666	.786	.883	.939	.987
7.0	s.p.	.117	.182	.255	.350	.482	.632	.781	.880	.974
	n.p.	.117	.234	.352	.478	.620	.751	.862	.927	.985
10	s.p.	.091	.152	.227	.327	.465	.621	.776	.877	.974
	n.p.	.091	.191	.300	.422	.569	.712	.838	.913	.982
16	s.p.	.067	.130	.209	.313	.456	.615	.773	.875	.973
	n.p.	.067	.149	.245	.361	.509	.662	.805	.895	.978
100	s.p.	.051	.117	.200	.307	.452	.613	.771	.875	.973
	n.p.	.051	.118	.201	.308	.454	.615	.773	.875	.973

Below the heavy line the s.p. test is less powerful then the n.p. test.

"power" of the "sometimes pool" test is greater everywhere except near $(\theta_{21}, \theta_{32}) = (1, 1)$.

3.2. Effect of P_2, P_3 illustrated. The influence of the probability levels employed in the final phase of a "sometimes pool" test is illustrated in Figure 3. The main effect is observed to be the manner in which the behaviour is constrained at the extreme values of θ_{21}.

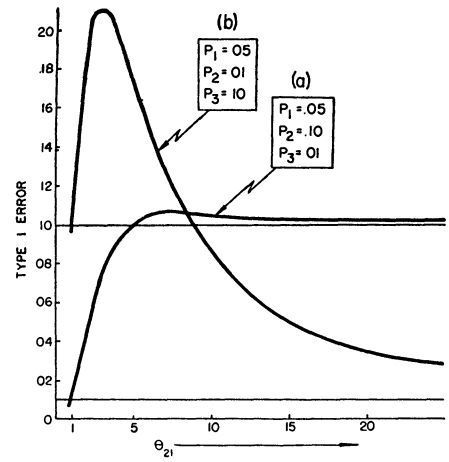

FIG. 3. Class A Tests; $n_1 = 20$, $n_2 = 4$, $n_3 = 2$.

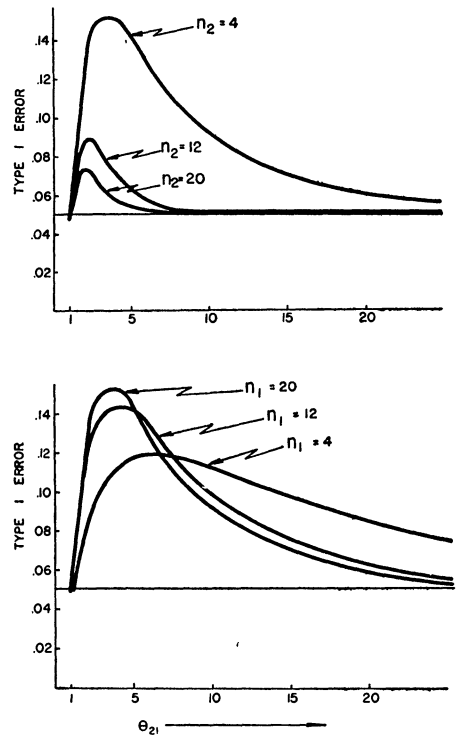

FIG. 4. (a) Upper Diagram: Effect of Varying n_2. $P_1 = P_2 = P_3 = .05$ and $n_1 = 20$, $n_3 = 2$. (b) Lower Diagram: Effect of Varying n_1. $P_1 = P_2 = P_3 = .05$ and $n_2 = 4$, $n_3 = 2$.

3.3. Effect of n_2, n_1 illustrated. The response of the Type I error to increases in the degrees of freedom of the preliminary test is illustrated in Figure 4. The maximum disturbance is observed to increase as n_1 increases or as n_2 decreases.

3.4. Class B test illustrated. The behaviour of the Type I error of some Class B tests is illustrated in Figure 1(b). The hypothesis is always rejected less frequently than if a "never pool" test at the P_2 level is used.

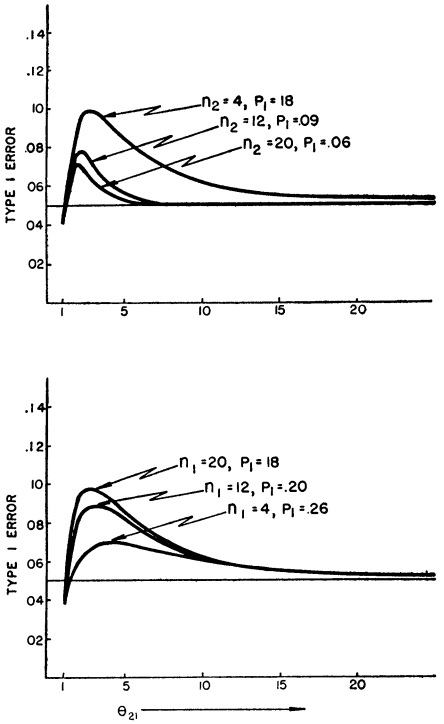

FIG. 5. (a) Upper Diagram: Effect of Varying n_2 when $F_1 = 2F_{50}$, $P_2 = P_3 = .05$ and $n_1 = 20$, $n_3 = 2$. (b) Lower Diagram: Effect of Varying n_1 when $F_1 = 2F_{50}$, $P_2 = P_3 = .05$ and $n_2 = 4$, $n_3 = 2$.

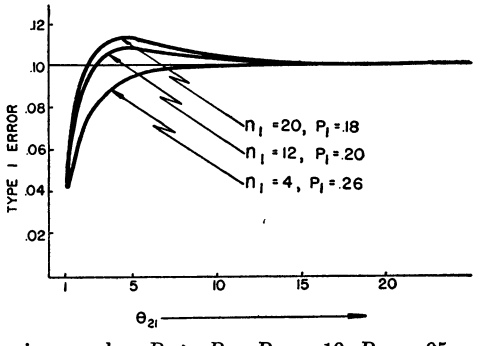

FIG. 6. Effect of Varying n_1 when $P_2 > P_3$. $P_2 = .10$, $P_3 = .05$ and $n_2 = 4$, $n_3 = 2$.

3.5. Recommended procedure illustrated. Figure 5 illustrates the behaviour of the Type I error when the recommended procedure is applied to the special cases presented in Figure 4. When $n_1 = 12$, $n_2 = 4$, the 20% probability level is

prescribed and the Type I error never exceeds .09. When $n_1 = 20$, $n_2 = 20$, the more liberal value of 6% is prescribed and the resulting Type I error never exceeds .07. The more liberal choice of P_1 results in a greater gain of power, near $\theta_{21} = 1$, than would have resulted if the 20% level had been used throughout. A small loss in power occurs when θ_{21} is large. Should the experimenter wish to guard against this loss in power for a larger range of values of θ_{21} near one, he may do so, at the expense of a somewhat larger disturbance in the Type I error, by choosing P_2 larger than P_3. In the present example, if P_2 is taken as .10 instead of .05, Figure 6 shows that the Type I error is changed only slightly for values of θ_{21} near one, but the maximum disturbance is increased. Such a test, is uniformly more powerful than the "never pool" test for all values of θ_{21} for which the Type I error is less than .10; a much larger range of values than in the previous case.

IV. Derivations and Proofs

4.1. Derivation of joint frequency function. The joint frequency function of the v's is given by

$$c_1 v_1^{\frac{1}{2}n_1-1} v_2^{\frac{1}{2}n_2-1} v_3^{\frac{1}{2}n_3-1} \exp\left\{-\tfrac{1}{2}\left[\frac{n_1 v_1}{\sigma_1^2} + \frac{n_2 v_2}{\sigma_2^2} + \frac{n_3 v_3}{\sigma_3^2}\right]\right\},$$

where c_1 is independent of the v's. Transform to new variables:

$$u_1 = \frac{n_2 v_2}{n_1 v_1}, \qquad u_2 = \frac{n_3 v_3}{n_2 v_2}, \qquad w = \frac{n_1 v_1}{n_3}.$$

By integrating and evaluating the constant, the joint frequency function of u_1 and u_2 is obtained:

$$p = \frac{\theta_{21}^{\frac{1}{2}n_1}\,\theta_{32}^{\frac{1}{2}(n_2+n_1)}}{\mathrm{B}(\frac{1}{2}n_2, \frac{1}{2}n_1)\mathrm{B}(\frac{1}{2}n_3, \frac{1}{2}(n_1+n_2))} \frac{u_1^{\frac{1}{2}(n_3+n_2)-1}\, u_2^{\frac{1}{2}n_3-1}}{(\theta_{21}\theta_{32} + \theta_{32}u_1 + u_1u_2)^{\frac{1}{2}(n_3+n_2+n_1)}} \tag{3}$$

where $\theta_{21} = \sigma_2^2/\sigma_1^2$; $\theta_{32} = \sigma_3^2/\sigma_2^2$.

4.2. Definition of critical region. The rule of procedure for the "sometimes pool" test may now be expressed in terms of the u's. Reject the hypothesis $\theta_{32} = 1$ if

$$\begin{cases} u_1 \geq u_1^0, \\ u_2 \geq u_2^0, \end{cases} \quad \text{or} \quad \begin{cases} u_1 < u_1^0, \\ \dfrac{u_1 u_2}{1+u_1} \geq u_3^0, \end{cases}$$

where

$$u_1^0 = \frac{n_2}{n_1} \cdot F_1(n_2, n_1 ; P_1),$$

$$u_2^0 = \frac{n_3}{n_2} \cdot F_2(n_3, n_2 ; P_2),$$

$$u_3^0 = \frac{n_3}{n_2+n_1} \cdot F_3(n_3, n_2+n_1 ; P_3).$$

The reader will note that the u's are ratios of sums of squares. The symbol u_1 is associated with the preliminary test. The final test when pooling is not prescribed is associated with the symbol u_2, and when pooling is prescribed the relevant statistic is $u_1u_2/(1 + u_1)$.

The critical region defined in this way is illustrated in the two dimensional sample space $\{u_1, u_2\}$ of Figure 7(a). The critical regions of the "never pool" and the "always pool" test are readily identified in this figure. The region of a "never pool" test at the P_2 level is designated by $A + B_1 + C$, the area above the line $u_2 = u_2^0$; and the region of an "always pool" test at the P_3 level is designated by $B_1 + B_2 + C + D$, the area above the curve $u_1u_2 = u_3^0(1 + u_1)$. The critical region of the "sometimes pool" test, $B_1 + B_2 + C$, may be considered in two parts: the portion due to *pooling*, $B_1 + B_2$, and the portion due to *not pooling*, C.

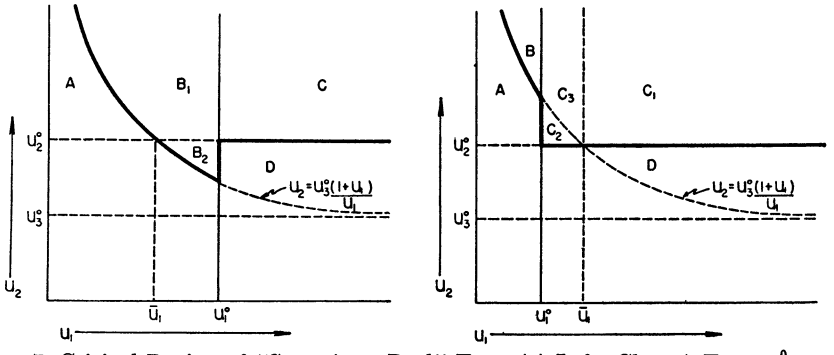

FIG. 7. Critical Region of "Sometimes Pool" Test. (a) Left: Class A Test: $u_1^0 > \bar{u}_1$ (b) Right: Class B Test: $u_1^0 < \bar{u}_1$.

The probability of rejecting the null hypothesis is given by

$$Q(\theta_{21}, \theta_{32}) = \int_0^{u_1^0} \int_w^\infty p \, du_1 \, du_2 + \int_{u_1^0}^\infty \int_{u_2^0}^\infty p \, du_1 \, du_2, \tag{4}$$

where p is the frequency function (3), and $w = u_3^0(1 + u_1)/u_1$.

Simple explicit expressions for these integrals cannot be obtained in general, but when $n_3 = 2$ they can be reduced to forms containing incomplete beta functions. This special case is dealt with in Subsection 4.7.

4.3. Critical value of P_1. The symbol $\bar{u}_1$ in Figure 1 is used to denote the u_1 coordinate of the point of intersection of the line $u_2 = u_2^0$ and the curve $u_1u_2 = u_3^0(1 + u_1)$. Accordingly,

$$\bar{u}_1 = \frac{u_3^0}{u_2^0 - u_3^0}, \tag{5}$$

a value readily determined for any given test. This relationship may be expressed in terms of the F's as

$$\bar{F}_1 = \frac{1}{\frac{n_2}{n_1}\left\{\frac{F_2}{F_3} - 1\right\} + \frac{F_2}{F_3}}, \tag{6}$$

where $\bar{F}_1$ is defined by $n_1\bar{u}_1 = n_2\bar{F}_1$. The probability level corresponding to $\bar{F}_1$ is denoted by $\bar{P}_1$.

The critical value $\bar{P}_1$ is the value of P_1 which divides the possible "sometimes pool" tests into two types having different properties. If P_1 is less than $\bar{P}_1(F_1 > \bar{F}_1$ or $u_1^0 > \bar{u}_1)$, the test is referred to as a Class A test. If P_1 is greater than $\bar{P}_1(F_1 < \bar{F}_1$ or $u_1^0 < \bar{u}_1)$, the test is referred to as a Class B test.

4.4. Lemma 1.

LEMMA 1. *If* $\theta'_{21} \geq \theta_{21}$ *and* $\theta'_{32} \geq \theta_{32}$, *and if the equality applies in one of these, then the ratio of the frequency functions* (3)

$$\frac{p(u_1, u_2 \mid \theta'_{21}, \theta'_{32})}{p(u_1, u_2 \mid \theta_{21}, \theta_{32})} \tag{7}$$

increases monotonically as (i) u_1 *increases with* u_2 *fixed, or as* (ii) u_2 *increases with* u_1 *fixed, or as* (iii) u_1 *increases on fixed pooling curve* $u_1u_2 = u_3^0(1 + u_1)$.

PROOF. The ratio (7) is a monotonic function of

$$\frac{\theta_{21}\theta_{32} + \theta_{32}u_1 + u_1u_2}{\theta'_{21}\theta'_{32} + \theta'_{32}u_1 + u_1u_2}.$$

It is easily shown that an expression of the form $(a + bx)/(c + dx)$ increases monotonically with respect to x if $a/c < b/d$, and this condition holds for cases (i), (ii), and (iii).

4.5. Lemma 2.

LEMMA 2. *If area* L *lies above a given pooling curve, and to the right of a given preliminary line, if area* K *lies below the same pooling curve, and to the left of the same preliminary line, and if*

$$Pr\{L \mid \theta_{21}, \theta_{32}\} \geq Pr\{K \mid \theta_{21}, \theta_{32}\},$$

then

$$Pr\{L \mid \theta'_{21}, \theta'_{32}\} > Pr\{K \mid \theta'_{21}, \theta'_{32}\},$$

where $\theta'_{21} \geq \theta_{21}$ *and* $\theta'_{32} \geq \theta_{32}$ *and the equality applies in one of these.*

PROOF. For any point (u_1, u_2) in K and any point (u'_1, u'_2) in L, Lemma 1 (iii) yields

$$\frac{p(u_1, u_2 \mid \theta'_{21}, \theta'_{32})}{p(u_1, u_2 \mid \theta_{21}, \theta_{32})} < \frac{p(u'_1, u''_2 \mid \theta'_{21}, \theta'_{32})}{p(u'_1, u''_2 \mid \theta_{21}, \theta_{32})},$$

where $u''_2 = c(1 + u'_1)/u'_1$, and c is a constant defined by $u_2 = c(1 + u_1)/u_1$. Since K is below a given pooling curve, $u''_2 < u'_2$ and

$$\frac{p(u'_1, u''_2 \mid \theta'_{21}, \theta'_{32})}{p(u'_1, u''_2 \mid \theta_{21}, \theta_{32})} < \frac{p(u'_1, u'_2 \mid \theta'_{21}, \theta'_{32})}{p(u'_1, u'_2 \mid \theta_{21}, \theta_{32})}.$$

Consider

$$\frac{p(u_1, u_2 \mid \theta'_{21}, \theta'_{32})}{p(u_1, u_2 \mid \theta_{21}, \theta_{32})} < b < \frac{p(u'_1, u'_2 \mid \theta'_{21}, \theta'_{32})}{p(u'_1, u'_2 \mid \theta_{21}, \theta_{32})},$$

where b is a constant such that the inequalities hold for all (u_1, u_2) in K and all (u'_1, u'_2) in L.

Integrating over the regions yields

$$Pr\{K \mid \theta'_{21}, \theta'_{32}\} < b.Pr\{K \mid \theta_{21}, \theta_{32}\}$$

and

$$b.Pr\{L \mid \theta_{21}, \theta_{32}\} < Pr\{L \mid \theta'_{21}, \theta'_{32}\}.$$

But

$$Pr\{K \mid \theta_{21}, \theta_{32}\} \leq Pr\{L \mid \theta_{21}, \theta_{32}\};$$

thus

$$Pr\{K \mid \theta'_{21}, \theta'_{32}\} < Pr\{L \mid \theta'_{21}, \theta'_{32}\},$$

which completes the proof.

4.6. General Properties.

RESULT 1. *When $\theta_{21} = 1$, the Type I error of a Class A test is less than P_3.*

PROOF. In the notation of Fig. 7(a), the probability of falling in $B_1 + B_2 + C + D$ is P_3 when $\theta_{21} = 1$ and $\theta_{32} = 1$. The region of rejection of the "sometimes pool" test is smaller by D.

RESULT 2. *When $\theta_{21} = 1$, the Type I error of a Class A test is greater than $(1 - P_1)P_3$.*

PROOF. The statistics u_1 and $u_1u_2/(1 + u_1)$ are independent when $\theta_{21} = 1$ and $\theta_{32} = 1$. Under these conditions, the probability of falling in $B_1 + B_2$, in the notation of Fig. 7(a), is equal to the product of two incomplete beta functions having the values $(1 - P_1)$ and P_3. Consequently, the Type I error is greater than $(1 - P_1)P_3$.

RESULT 3. *The Type I error approaches P_2 as θ_{21} approaches infinity.*

PROOF. The distribution becomes singular when $\theta_{21} = \infty$. The frequency function approaches zero uniformly for any finite value of u_1 and approaches

$$\frac{1}{B(\frac{1}{2}n_3, \frac{1}{2}n_2)} \frac{u_2^{\frac{1}{2}n_3 - 1}}{(1 + u_2)^{\frac{1}{2}(n_3 + n_2)}}$$

at $u_1 = \infty$. When $\theta_{21} = \infty$, the entire mass is concentrated on the line $u_1 = \infty$ and is distributed as a beta variable along that line. In the notation of Fig. 7(a), $Pr(B_1 + B_2) \to 0$ and $Pr(C) \to P_2$.

RESULT 4. *If the Type I error of a Class A test is Q_0 for θ_{21}, then for $\theta'_{21} > \theta_{21}$, the Type I error is greater than r, where r is equal to the lesser of Q_0 and P_2.*

Three useful corollaries are associated with the above result:

RESULT 4.1. *If at* $\theta_{21} = 1$, *the value of the Type I error is less than* P_2, *this is its minimum value for any* θ_{21}.

RESULT 4.2. *If at* $\theta_{21} = 1$, *the Type I error is less than* P_2, *then as* θ_{21} *increases from* 1 *the Type I error increases monotonically until* P_2 *is reached.*

RESULT 4.3. *If for some value of* θ_{21} *the Type I error is equal to or greater than* P_2, *then for any larger value of* θ_{21}, *the Type I error is greater than* P_2.

PROOF. Let the regions of Fig. 8 be denoted by $R_1 = A_1 + B_1 + C_1$ with similar designations for R_2 and R_3. Let $R_4 = B_1 + B_2 + B_3 + B_4 + C_1 + C_2$.

If $r = Q_0$, let the non-pooling line between R_1 and R_2 in Fig. 8 correspond to Q_0 for all θ_{21}. Then $Pr\{R_4 \mid \theta_{21}, 1\} = Pr\{R_1 \mid \theta_{21}, 1\}$, whence $Pr\{B_2 + B_3 + B_4 + C_2 \mid \theta_{21}, 1\} = Pr\{A_1 \mid \theta_{21}, 1\}$. By Lemma 2, we have for any $\theta'_{21} > \theta_{21}$, $Pr\{B_2 + B_3 + B_4 + C_2 \mid \theta'_{21}, 1\} > Pr\{A_1 \mid \theta'_{21}, 1\}$ and $Pr\{R_4 \mid \theta'_{21}, 1\} > Pr\{R_1 \mid \theta'_{21}, 1\} = Q_0$.

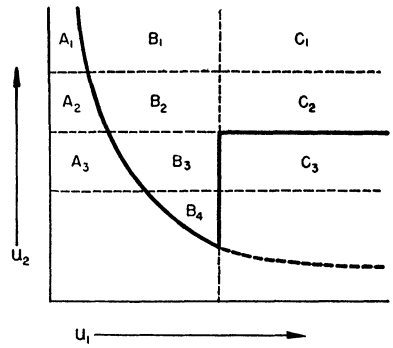

FIG. 8. Critical Regions for Result 4.

If $r = P_2$, let the non-pooling line at the lower boundary of R_3 in Fig. 8 correspond to Q_0 for all θ_{21}. Then in the same way $Pr\{B_4 \mid \theta_{21}, 1\} = Pr\{A_1 + A_2 + A_3 + C_3 \mid \theta_{21}, 1\}$ and $Pr\{B_4 \mid \theta'_{21}, 1\} > Pr\{A_1 + A_2 + A_3 \mid \theta'_{21}, 1\}$ by Lemma 2. Thus $Pr\{R_4 \mid \theta'_{21}, 1\} > Pr\{R_1 + R_2 + A_3 + B_3 \mid \theta'_{21}, 1\}$ and $Pr\{R_4 \mid \theta'_{21}, 1\} > Pr\{R_1 + R_2 \mid \theta'_{21}, 1\} = P_2$.

RESULT 5. *For a Class B test, the Type I error is less than* P_2 *for all* θ_{21}.

PROOF. Figure 7(b) illustrates the critical region of a Class B test. We have $Pr\{A + B + C_1 + C_2 + C_3\} = P_2$. But the region of rejection of the "sometimes pool" test is smaller, excluding A.

RESULT 6. *The Type I error of a Class B test, for* $\theta_{21} = 1$, *is greater than* $(1 - \bar{P}_1)P_3$.

PROOF. Changing P_1 to $\bar{P}_1$ removes C_2 from the region of rejection in Fig. 7(b), thus decreasing the Type I error. The modified test lies in both Class B and Class A, so that Result 2 applies.

RESULT 7. *For any* θ_{21}, *the Type I error is a minimum for changes of* P_1 *when* $P_1 = \bar{P}_1$.

PROOF. For a Class A test, changing P_1 to $\bar{P}_1$ removes region B_2 of Fig. 7(a), thus decreasing the Type I error. For a Class B test, changing P_1 to $\bar{P}_1$ removes region C_2 of Fig. 7(b), similarly decreasing the Type I error.

RESULT 8. *A Class A test, in which the Type I error is less than or equal to P_2, is more powerful than a "never pool" test having the same Type I error.*

PROOF. In Fig. 8, let region $R_1 = A_1 + B_1 + C_1$ be equal in size to $R_4 = B_1 + B_2 + B_3 + B_4 + C_1 + C_2$. Then $Pr\{R_4 \mid \theta_{21}, 1\} = Pr\{R_1 \mid \theta_{21}, 1\}$ and $Pr\{B_2 + B_3 + B_4 + C_2 \mid \theta_{21}, 1\} = Pr\{A_1 \mid \theta_{21}, 1\}$. Increasing $\theta_{32} = 1$ to θ'_{32} and applying Lemma 2 yields $Pr\{R_4 \mid \theta_{21}, \theta'_{32}\} > Pr\{R_1 \mid \theta_{21}, \theta_{32}\}$.

RESULT 9. *For a fixed Type I error a Class A test, carried out at given levels of P_2 and P_3, is more powerful than a Class B test at the same levels.*

PROOF. Fig. 7 and Lemma 2 apply at once.

4.7. Closed form expressions for $n_3 = 2$. The probability of rejecting the hypothesis in a "sometimes pool" test is given by $Q(\theta_{21}, \theta_{32}) = Q_1 + Q_2$ where Q_1 corresponds to the region B, and Q_2 to the region C of Fig. 7.

The integrals (4) representing the probability of rejecting the null hypothesis, reduce, when $n_3 = 2$, to

$$Q_1 = \left\{\frac{1 + \dfrac{u_3^0}{\theta_{32}}}{1 + \dfrac{u_3^0}{\theta_{21}\theta_{32}}}\right\}^{\frac{1}{2}n_1} \frac{I_z(\frac{1}{2}n_2, \frac{1}{2}n_1)}{\left\{1 + \dfrac{u_3^0}{\theta_{32}}\right\}^{\frac{1}{2}(n_2+n_1)}}, \tag{8}$$

where the argument z of the incomplete beta function is defined by $z = x/(1 + x)$ where

$$x = \left\{\frac{1 + \dfrac{u_3^0}{\theta_{32}}}{1 + \dfrac{u_3^0}{\theta_{21}\theta_{32}}}\right\}. \tag{9}$$

Under the null hypothesis $\theta_{32} = 1$,

$$Q_1 = I_z(\tfrac{1}{2}n_2, \tfrac{1}{2}n_1)\left\{\frac{1 + u_3^0}{1 + \dfrac{u_3^0}{\theta_{21}}}\right\}^{\frac{1}{2}n_1} \cdot P_3, \tag{10}$$

since

$$P_3 = \frac{1}{(1 + u_3^0)^{\frac{1}{2}(n_2+n_1)}}.$$

Similarly

$$Q_2 = \frac{I_{z'}(\frac{1}{2}n_1, \frac{1}{2}n_2)}{\left\{1 + \dfrac{u_2^0}{\theta_{32}}\right\}^{\frac{1}{2}n_2}}, \tag{11}$$

where the argument z' of the incomplete beta function is defined by $z' = 1/(1+x')$ where

$$x' = \left\{1 + \frac{u_2^0}{\theta_{32}}\right\} \frac{u_1^0}{\theta_{21}}. \tag{12}$$

Under the null hypothesis $\theta_{32} = 1$,

$$Q_2 = I_{z'}(\tfrac{1}{2}n_1 , \tfrac{1}{2}n_2) \cdot P_2 , \tag{13}$$

since

$$P_2 = \frac{1}{(1 + u_2^0)^{\frac{1}{2}n_2}}.$$

The incomplete beta function has been tabulated by Pearson [4].

The author wishes to thank Professor W. G. Cochran and Professor John W. Tukey for helpful advice in the preparation of this paper.

REFERENCES

[1] T. A. Bancroft. "On biases in estimation due to the use of preliminary tests of significance". *Annals of Math. Stat.*, Vol. 15 (1944), pp. 190–204.

[2] Frederick Mosteller. "On pooling data". *Jour. Am. Stat. Assn.*, Vol. 43 (1948), pp. 231–242.

[3] M. Merrington and C. M. Thompson. "Tables of percentage points of the inverted beta (F) distribution". *Biometrika*, Vol. 33 (1943), pp. 73–88.

[4] Karl Pearson, *Tables of the Incomplete Beta Function*, Cambridge University Press, 1934.