Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 8183–8184, November 1989
Statistics

One-way analysis of variance with unequal variances

(data analysis/Behrens–Fisher problem/biostatistics)

WILLIAM R. RICE[†‡] AND STEVEN D. GAINES[§]

[†]Department of Biology, University of New Mexico, Albuquerque, NM 87131; and [§]Program in Ecology and Evolutionary Biology, Box G, Brown University, Providence, RI 02912

Communicated by John Imbrie, August 7, 1989 (received for review February 1, 1989)

ABSTRACT **We have designed a statistical test that eliminates the assumption of equal group variances from one-way analysis of variance. This test is preferable to the standard technique of trial-and-error transformation and can be shown to be an extension of the Behrens–Fisher *T* test to the case of three or more means. We suggest that this procedure be used in most applications where the one-way analysis of variance has traditionally been applied to biological data.**

Fixed effects one-way analysis of variance I (ANOVA-1W) is a frequently used statistical tool in many areas of science, especially the biological sciences. A difficulty inherent in its application, however, is the common occurrence of inter-group heteroscedasticity—i.e., variances differ among groups. The ANOVA-1W test requires that variances be homogeneous.

One solution to the heteroscedasticity problem is based upon sequential statistical analysis (1). This technique requires that two samples be taken from each population, with the sizes of the second samples contingent upon the characteristics of the first samples. Such conditional, multiple sampling is not practical in many biological settings, and a nonsequential procedure is warranted.

Transformation is a nonsequential technique that may be used to correct for heteroscedasticity, but exact significance testing is only possible when a suitable transformation can be prescribed *a priori*. An approximate test is possible when trial and error leads to a transformation that, based upon residual analysis, eliminates heteroscedasticity while maintaining normality of group means. Even this approximate test is inappropriate when sample sizes are small, however, because residual analysis cannot reliably determine the suitability of a transformation. A second complicating factor is that elimination of heteroscedasticity by transformation tends to skew the transformed data and thereby violates the assumption of the normality of sample means. Exact, nonsequential testing in the context of ANOVA-1W is therefore rarely possible when heteroscedasticity is encountered, and a researcher is left with the dilemma of no suitable parametric test.

Here we solve this problem by developing an ANOVA-1W procedure that requires neither equality of the group variances nor sequential sampling. We begin by first expressing Fisher's F statistic (2), from the ANOVA-1W, in terms of Student's t tests between all pairwise combinations of means. Fisher's F statistic from an ANOVA-1W is typically expressed as the ratio of the between-group to the within-group mean-squared errors—i.e.,

$$F = \{SSB/(K-1)\}/\{SSE/(N_T - K)\} \quad [1]$$

$$= [1/(K-1)] \sum (Y_{.i} - Y_{..})^2/(S_p^2/N_i),\ i = 1, \ldots, K, \quad [2]$$

where SSB and SSE are the between-group and within-group sums of squares, respectively, K is the number of groups, $N_T = N_1 + N_2 + \ldots + N_K$, where N_i is the sample size of the ith group, S_p^2 is the pooled variance across all groups ($S_p^2 = SSE/[N_T - K]$), $Y_{.i}$ is the sample mean for the ith group, and $Y_{..}$ is the weighted average grand mean with weights equaling $1/(S_p^2/N_i)$.

An algebraically equivalent expression is

$$F = \text{weighted average } (t_{ij}^2);\quad i = 1, \ldots, (K-1),\ j = (i+1), \ldots, K, \quad [3]$$

where t_{ij}^2 is the squared t value from a Student's t test comparing the ith and jth means ($t_{ij} = \{Y_{.i} - Y_{.j}\}/\{S_p^2[1/N_i + 1/N_j]\}^{1/2}$), and $weight_{ij} = 1/(S_p^2/N_i) + 1/(S_p^2/N_j)$. The P value for the F statistic can therefore be interpreted as the probability, on the null hypothesis (H_0; $u_1 = u_2 = \ldots = u_k$) with repeated sampling, that the average t^2 statistic will be greater than or equal to the observed average. Because a test based upon the square of a t statistic is equivalent to a two-tailed t test, Fisher's F statistic from the ANOVA-1W can be interpreted as the average two-tailed t test between means.

To develop a test statistic when equal variances cannot be presumed, we begin by first considering the simplest case in which the true variances for each group are known without error. A new test statistic, F^*, can be defined in terms of the normal Z tests between all pairwise combinations of means,

$$F^* = \text{weighted average } (Z_{ij}^2);\quad i = 1, \ldots, (K-1),\ j = (i+1), \ldots, K \quad [4]$$

$$= [1/(K-1)] \sum (Y_{.i} - Y'_{..})^2/(V_{.i}^2/N_i);\ i = 1, \ldots, K, \quad [5]$$

where $Y_{.i}$, N_i, and K are defined as before, Z_{ij}^2 is the squared Z value from a normal Z test comparing the ith and jth means ($Z_{ij} = \{Y_{.i} - Y_{.j}\}/\{[V_i^2/N_i] + [V_j^2/N_j]\}^{1/2}$), $weight_{ij} = 1/(V_{.i}^2/N_i) + 1/(V_{.j}^2/N_j)$, $V_{.i}^2$ is the known variance for the ith group, $Y'_{..}$ is the weighted average grand mean with weights equaling $1/(V_i^2/N_i)$. The F^* statistic follows a χ^2 divided by degrees of freedom distribution, with degrees of freedom equaling $K - 1$. Small values of F^* support H_0, whereas large values support the H_A that $u_i \neq u_j$ for at least one i,j pair. The P value from a test based upon the F^* statistic is the probability, on H_0 with repeated sampling, that F^* will be greater than or equal to the observed value. This is the $\text{prob}(X_{K-1}^2/(K-1) \geq F^*_{\text{observed}})$ in repeated sampling, where X_{K-1}^2 is a χ^2 variate with $K - 1$ degrees of freedom.

When the variances of the groups are unknown, the F^* statistic is inappropriate because the Z_{ij} cannot be calculated for the pairwise tests between means. As shown above, however, replacing the Z_{ij} with Student's t values leads to

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: ANOVA-1W, fixed effects one-way analysis of variance (that assumes equal variances).

[‡]To whom reprint requests should be addressed.

Fisher's F statistic, which enables testing when there is a common but unknown variance among groups. When the variances are unknown and differ among groups, we propose replacing the Z_{ij} with Behrens–Fisher T values ($T_{ij} = \{Y_{.i} - Y_{.j}\}/\{S_i^2/N_i + S_j^2/N_j\}^{1/2}$). Barnard (3) has recently discussed the appropriateness of the Behrens–Fisher T test (2, 4) when comparing means derived from normal distributions with unequal variances. He concludes that P values from the Behrens–Fisher T test need not be motivated by fiducial arguments and that T is the preferred nonsequential test statistic for comparing means from normal distributions with unequal and unknown variances.

To incorporate the Behrens–Fisher T statistic into ANOVA-1W we define the modified F^* statistic, F_U, in which the V_i^2 are replaced by their sample estimates,

$$F_U = [1/(K-1)] \sum (Y_{.i} - Y''_{..})^2/(S_i^2/N_i), \quad i = 1, \ldots, K \quad [6]$$

$$= \text{weighted average } (T_{ij}^2); \quad i = 1, \ldots, (K-1), \quad j = (i+1), \ldots, K, \quad [7]$$

where $Y_{.i}$, N_i, and K are defined as before, S_i^2 is the nonpooled sample variance for the ith mean, $Y''_{..}$ is the weighted average grand mean with weights equaling $1/(S_i^2/N_i)$, T_{ij} is the test statistic from the Behrens–Fisher test between means i and j, and $weight_{ij} = 1/(S_i^2/N_i) + 1/(S_j^2/N_j)$. Note that: (*i*) The F and F_U statistics are special cases of F^* in which the sample variances are substituted for the population variances, (*ii*) for the case of two means the F_U statistic reduces to the square of the Behrens–Fisher T statistic (i.e., $F_U = T^2$), and (*iii*) F^* is a weighted-average Z^2, F_U is a weighted-average T^2, and F is a weighted average t^2.

For testing purposes, one rejects the null hypothesis H_0 in favor of the alternate hypothesis H_A for large values of F_U. The P value from a test based upon F_U is the probability, on H_0, that $F_U \geq F_{U(\text{observed})}$. As was the case for Fisher's F, the F_U statistic can be interpreted as an average two-tailed test between means.

The distribution of F_U converges on that of a χ^2 variate divided by degrees of freedom ($X_{K-1}^2/[K-1]$) when all sample sizes jointly approach infinity. The cumulative distribution function of F_U will be described in detail elsewhere and can be shown to equal,

$$\text{Prob}(F_U \geq F_{U(\text{observed})}) = \int_0^\infty \cdots \int_0^\infty \text{Prob}_\chi(f_U) \cdot L(v_1^2|S_1^2) \cdot \ldots \cdot L(v_k^2|S_k^2) dv_1^2 \ldots dv_k^2, \quad [8]$$

where

$$f_U = [1/(K-1)] \sum (Y_{.i} - y''_{..})^2/(v_i^2/N_i), \quad i = 1, \ldots, K, \quad [9]$$

Prob_χ is the probability that ($X_{K-1}^2/[K-1] \geq f_U$), $L(v_i^2|S_i^2)$ is the likelihood of v_i^2 given its sample estimate (S_i^2), and $y''_{..}$ is the weighted average grand mean with weights equaling $1/(v_i^2/N_i)$.

Tabulation of the critical values for F_U is not practical, due to the large number of parameters (i.e., $k, n_1, \ldots, n_k, S_1^2, \ldots, S_k^2$) associated with the F_U distribution, nor is it necessary. We have developed a simple numerical procedure, based on a technique developed by Barnard (3), for calculating exact P values for observed values of F_U—i.e., the Prob($F_U \geq F_{U(\text{observed})}$) when H_0 is true. These calculations can be done by hand but become tedious for large values of K—a problem we have solved with a simple computer program. The numerical procedure and its computerized solution will be described elsewhere as part of a more detailed comparison of the power and robustness to deviations from model assumptions of the F_U test.

As a numerical example of the F_U test, suppose three populations were sampled and that $N_i = 5, 6, 7$; $S_i^2 = 4, 7, 25$; and $Y_{.i} = 0, 5, 2$ for samples 1, 2, and 3, respectively. A Bartlett test (5) with these data leads to acceptance of the null hypothesis of equal group variances ($P = 0.137$). If the potential heteroscedasticity were ignored, the F statistic from an ANOVA-1W would be 2.629, and the corresponding P value would be 0.105. When our test is applied, $F_U = 6.356$, and the corresponding P value is 0.036. This example illustrates how the P values from the new procedure can be considerably smaller than those from an ANOVA-1W. This is a consequence of the procedure described here not artificially inflating the smaller variance estimates, as would occur via the pooling procedure used in ANOVA-1W.

We suggest that the F_U test be used in those cases where the unknown group variances cannot be demonstrated to be equal. This situation will include most cases where ANOVA-1W has traditionally been used, either with or without transformation because the presumption of no intergroup heteroscedasticity can rarely be rigorously demonstrated.

This research was supported by National Science Foundation Grant BSR 8407440 (W.R.R.) and Department of Energy Contract EV10108 and Brown University Research Support Grant (S.D.G.).

1. Bishop, T. A. & Dudewicz, E. J. (1978) *Technometrics* **20**, 419–430.
2. Fisher, R. A. (1959) *Statistical Methods and Scientific Inference* (Oliver & Boyd, Edinburgh), 2nd Ed.
3. Barnard, G. A. (1984) *Appl. Stat.* **33**, 266–271.
4. Behrens, W. U. (1964) *Biometrics* **20**, 16–27.
5. Neter, J., Wasserman, W. & Kutner, M. H. (1985) *Applied Linear Statistical Models* (Irwin, Homewood), 2nd Ed.