$$\mathbf{X}_{tj} = \boldsymbol{\mu} + \boldsymbol{\alpha}_t + \mathbf{E}_{tj},$$

$(t = 1, \ldots, k;\ j = 1, \ldots, n_t)$, where $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}_t$ are each constant $p \times 1$ vectors, such that

$$\sum_{t=1}^{k} n_t \boldsymbol{\alpha}_t = \mathbf{0},$$

and $\mathbf{E}_{tj}$ is a random $p \times 1$ error vector. The joint distribution of the p-variates of $\mathbf{E}_{tj}$ does not depend on j and the expected value of each is 0. The common distribution is often assumed to be multinormal*, and the $\mathbf{E}_{tj}$'s are assumed mutually independent.

In multivariate analysis of variance* (MANOVA), the null hypothesis (of no difference among group means)

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \cdots = \boldsymbol{\alpha}_k = \mathbf{0} \qquad (1)$$

is tested against the class of alternative hypotheses that at least one of the equalities in (1) is violated.

For a general model with $p > 1$ and $k > 2$ and with the variance-covariance matrices of the distributions of the $\mathbf{E}_{tj}$'s unknown, James [1] proposed the test statistic

$$T_v^2 = \sum_{t=1}^{k} (\overline{\mathbf{X}}_t - \overline{\mathbf{X}})' \mathbf{W}_t (\overline{\mathbf{X}}_t - \overline{\mathbf{X}}),$$

where

$$\overline{\mathbf{X}}_t = n_t^{-1} \sum_{j=1}^{n_t} \mathbf{X}_{tj},$$

$$\mathbf{S}_t = (n_t - 1)^{-1} \sum_{j=1}^{n_t} (\mathbf{X}_{tj} - \overline{\mathbf{X}}_t)(\mathbf{X}_{tj} - \overline{\mathbf{X}}_t)',$$

$$\mathbf{W}_t = (n_t^{-1} \mathbf{S}_t)^{-1}, \quad \mathbf{W} = \sum_{t=1}^{k} \mathbf{W}_t,$$

$$\overline{\mathbf{X}} = \mathbf{W}^{-1} \sum_{t=1}^{k} \mathbf{W}_t \overline{\mathbf{X}}_t.$$

See related entries and James [1] for further details.

REFERENCE

1. James, G. S. (1954). *Biometrika*, **41**, 19–43.

See also BEHRENS–FISHER PROBLEM; MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA); WALD'S *W*-STATISTICS; and WELCH TESTS.

WELCH TESTS

In the one-way* layout to compare the means of k normally distributed populatons, it may not be valid in some cases to assume homogeneous variances. Hence the ANOVA* F-test* is not applicable, and the Welch [19] test was proposed to fill this void. An important special case ($k = 2$) is the famous Behrens–Fisher* problem. This special case was solved by Welch [18] several years earlier than the general case. His solution for $k = 2$ was refined and tabled by Aspin [1,2] and has become known as the Aspin–Welch test (AWT). Further tables were later provided by Trickett et al. [17]. Competing solutons to the Behrens–Fisher problem have been suggested by Fisher [8], Lee and Gurland [11] (denoted LG), Cochran [6], and Welch himself [18; 2, Appendix]. All these tests depend on normality, and Yuen [21] and Tiku and Singh [16] attempt more robust solutions. Some competing procedures for general k are due to Brown and Forsythe [5], James [9], and Bishop and Dudewicz [3]. The Welch and Brown-Forsythe tests have been extended by Roth [13] to the case where the k populations have a natural ordering (e.g., different dosages of the same drug) and a trend test* is desired to detect differences in the means that are monotone as a function of this ordering.

Another (unrelated) Welch [20] test was designed in mixed or random effects models to provide confidence intervals for variance components*, whose estimators are often distributed as linear combinations of chi-squared variates. Basically, Welch provides correction terms to the confidence limits obtained via the Satterthwaite [14,15] approximation*, which is based on a single chi-squared variate. These corrections were long among the most widely advocated methods (see, e.g., Mendenhall [12, pp. 352–354]), with no perceived major drawbacks except tedious computations. However, Boardman's [4] simulations showed that the Welch corrections are actually detrimental

to achieving the nominal confidence coefficients. Hence they have fallen justifiably into disfavor and will not be discussed further.

We now explore the details and properties of the Welch tests described in the first paragraph, and we begin with the AWT. The test statistic is

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \tag{1}$$

and Welch [18], for a slightly more general problem, derives the percentage points of t' as a power series in $1/f_i = 1/(n_i - 1)$ for $i = 1, 2$. The P fractile $(0 < P < 1)$ of t' is explicitly given to order $(1/f_i)^2$ by

$$\alpha\left[1 + \frac{(1+\alpha)^2}{4}\frac{\sum(s_i^4/n_i^2 f_i)}{(\sum s_i^2/n_i)^2} + \frac{(3+5\alpha^2+\alpha^4)}{3}\frac{\sum(s_i^6/n_i^3 f_i^2)}{(\sum s_i^2/n_i)^3} - \frac{(15+32\alpha^2+9\alpha^4)}{32}\frac{\sum(s_i^4/n_i^2 f_i)^2}{(\sum s_i^2/n_i)^4}\right], \tag{2}$$

where $\alpha = \Phi^{-1}(P)$ and Φ is the standard normal CDF. Note that the constant term in (2) reflects simply the normal approximation to t'.

Welch also suggests a method that refers to ordinary t-tables. This is done by equating the first two moments of t' to those of a t-distribution* with f degrees of freedom. The solution for f is

$$\frac{1}{f} = \frac{c^2}{f_1} + \frac{(1-c)^2}{f_2}, \tag{3}$$

where $c = (s_1^2/n_1)/(s_1^2/n_1 + s_2^2/n_2)$. Welch [18] originally suggested replacing f_i in (3) by $f_i + 2, i = 1, 2$, and blank, but he later repudiated this suggestion [2, Appendix]. He showed that critical values based on (3) agree with the correct ones based on (2) to order $(1/f_i)$, but they differ in the $(1/f_i)^2$ term.

The AWT is more powerful (i.e., has lower critical values) to varying degrees, and hence gives narrower confidence intervals, in general, than the essentially Bayesian* Behrens–Fisher solution, the Cochran method, or the Welch test based on (3); the latter two are widely used due to the computational simplicity of referring to ordinary t-tables, (3) being far more accurate. This accuracy is evaluated from the tables of LG [11], whose general method for this whole class of size and power calculations revealed that the AWT operates closest by far to the nominal level from among a set of seven competing tests. LG then proposed their own test, which is almost identical to the AWT in both size and power, and recommended it on the grounds of greater simplicity. However, it requires five constants that depend on the sample sizes and the nominal level, which are provided only for $5 \leqslant n_1 \leqslant n_2 \leqslant 10$ at the 0.05 level.

The Welch [19] test for general k compares the statistic

$$W^* = \frac{\sum w_j(\bar{x}_j - \hat{\mu})^2/(k-1)}{1 + [2(k-2)/(k^2-1)]\sum h_j}, \tag{4}$$

to the $F(k-1, f)$ distribution, where

$$w_j = n_j/s_j^2, \quad \hat{\mu} = \sum w_j x_j/W, \quad W = \sum w_j,$$
$$h_j = (1 - w_j/W)^2/(n_j - 1),$$
$$f = (k^2 - 1)/(3\sum h_j).$$

It and the Brown–Forsythe [5] test both reduce to the Welch test based on (3) when $k = 2$. The derivation of W^*, like that of the AWT, stems from a power series in $(1/f_i)$. Welch shows that W^* agrees to order $1/f_i$, but not to order $(1/f_i)^2$, with the James [9] test, which is based on a chi-squared (not F) approximation. Brown and Forsythe [5] demonstrate via simulations that, in general, their procedure and W^* both outperform the James test; furthermore, W^* tends to be better than their procedure when extreme means are associated with small variances, and vice versa. Importantly, both procedures lose little power in the equal variance case relative to the "optimal" ANOVA F-test, which is hence NOT recommended for the one-way layout. Dykstra and Werter [7] refine the James test and claim from their simulations that this refinement is on balance superior to the other tests; however, their numerical tables seem to support this conclusion only mildly

when $k = 6$ and not at all when $k = 4$. In any case, the Welch test is quite competitive. Incidentally, Johansen [10] rederives the Welch test as a special case of a more general result on residuals* from a weighted linear regression*.

Roth's [13] extension of W^* to the Welch trend test (WT) for ordered populations is basically obtained by first amalgamating the population means using isotonic regression* for simple order with weights $w_j = n_j/s_j^2$. Conditionally on the results of the amalgamation process, the statistic W^* (when applied to the amalgamated populations) is multiplied by an appropriate constant so that its conditional distribution is similar to that of $\overline{E}^2$, which is the trend analog of the ANOVA F-test. Roth also developed the Brown–Forsythe trend test (BFT), and his simulations showed that WT is generally (but by no means uniformly) the better of the two, tending to have larger type I error rates but compensating for this with gains in power too great to be explained merely by the differences in level. Conditions under which WT is superior to BFT (and vice versa) are analogous to the above-mentioned findings of Brown and Forsythe in the nontrend situation. Analogously as well, the $\overline{E}^2$-test does not seem to gain much power (and hence is not recommended) even when the variances are equal, unless the sample sizes are as small as 2 or 3.

REFERENCES

1. Aspin, A. A. (1948). *Biometrika*, **35**, 88–96. (Refines the AWT and tables some critical values.)
2. Aspin, A. A. (1949). *Biometrika*, **36**, 290–296. [Contains an appendix by Welch commenting on Aspin's work and proposing the test based on (3).]
3. Bishop, T. A. and Dudewicz, E. J. (1978). *Technometrics*, **20**, 419–430.
4. Boardman, T. J. (1974). *Biometrics*, **30**, 251–262. (Simulations and references involving 12 procedures for confidence intervals for variance components.)
5. Brown, M. B. and Forsythe, A. B. (1974). *Technometrics*, **16**, 385–389.
6. Cochran, W. G. (1964). *Biometrics*, **20**, 191–195. (Points out important drawbacks to his own procedure.)
7. Dykstra, J. B. and Werter, P. S. P. J. (1981). *Commun. Statist. B*, **10**, 557–569.
8. Fisher, R. A. (1941). *Ann. Eugen., Lond.*, **11**, 141–172. [The original fiducial (essentially Bayesian) solution to the Behrens–Fisher problem.]
9. James, G. S. (1951). *Biometrika*, **38**, 324–329.
10. Johansen, S. (1980). *Biometrika*, **67**, 85–92.
11. Lee, A. F. S. and Gurland, J. (1975). *J. Amer. Statist. Ass.*, **70**, 933–941. [Extensive bibliography on the entire subject. Also outlines general method for computing size and power of Welch-type tests (numerically tabled for several tests), and proposes a new test for $k = 2$.]
12. Mendenhall, W. (1968). *Introduction to Linear Models and the Design and Analysis of Experiments*. Wadsworth, Belmont, CA.
13. Roth, A. J. (1983). *J. Amer. Statist. Ass.*, **78**, 972–980. (Defines WT and BFT with detailed numerical examples.)
14. Satterthwaite, F. E. (1941). *Psychometrika*, **6**, 309–316.
15. Satterthwaite, F. E. (1946). *Biometrics Bull.*, **2**, 110–114.
16. Tiku, M. L. and Singh, M. (1981). *Commun. Statist. A*, **10**, 2057–2071. (Robust statistic for $k = 2$ that outperforms Yuen's.)
17. Trickett, W. H., Welch, B. L., and James, G. S. (1956). *Biometrika*, **43**, 203–205. (More tables of critical values for the AWT.)
18. Welch, B. L. (1947). *Biometrika*, **34**, 28–35. [Proposes both the AWT and the test based on (3) with f_i replaced by $f_i + 2$.]
19. Welch, B. L. (1951). *Biometrika*, **38**, 330–336. (Welch's k-sample test.)
20. Welch, B. L. (1956). *J. Amer. Statist. Ass.*, **51**, 132–148.
21. Yuen, K. K. (1974). *Biometrika*, **61**, 165–170. [Obtains level of Welch test based on (3) for many nonnormal distributions and proposes a more robust statistic.]

Editorial Note

In more recent papers, Aucamp (1986) (*J. Statist. Comp. Simul.*, **24**, 33–46) proposes the critical region

$$|t'| > z_{1-\alpha/2}[1 + 2\hat{C}^2 f_1^{-1} + 2(1 - \hat{C})^2 f_2^{-1}]^{1/2},$$

with $\Phi(z_{1-\alpha/2}) = 1 - \alpha/2$ and

$$\hat{C} = \frac{s_1^2}{n_1}\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^{-1},$$

and Matuszewski and Sotres (1986) (*Comp. Statist. Data Anal.*, **3**, 241–249) propose rejection of the null hypothesis if the 80% confidence intervals for the two individual means do not overlap—giving a significance level of approximately 5%.

See also BEHRENS–FISHER PROBLEM; ISOTONIC INFERENCE; TREND; and TREND TESTS.

ARTHUR J. ROTH

WELDON, WALTER FRANK RAPHAEL

Born: Highgate, London, England, March 15, 1860.

Died: Oxford, England, April 13, 1906.

Contributed to: Biometrics, evolutionary biology, zoology.

W. F. R. Weldon was the second child of the journalist and industrial chemist Walter Weldon and his wife Anne Cotton. His father changed residences so frequently that Weldon's early education was desultory until he became a boarder in 1873 at Caversham near Reading. Weldon matriculated at University College London (UCL) in the autumn of 1876 with the intention of pursuing a medical career. During his time at UCL, he acquired a respectable knowledge of mathematics from the Danish mathematician Olaus Henrici, and attended the lectures of the zoologist E. Ray Lankester. In the following year he transferred to Kings College, London, and in April 1878 he entered St. John's College, Cambridge, as a bye-term student.

Once at Cambridge, he met the zoologist Francis Maitland Balfour, and subsequently gave up his medical studies for zoology. In 1881, he gained a first-class degree in the Natural Science Tripos; in the autumn he left for the Naples Zoological Station to begin the first of his studies in marine biological organisms.

Upon returning to Cambridge in 1882, Weldon was appointed university lecturer in invertebrate morphology. In the following year he married Florence Tebb. He became a founding member of the Marine Biological Station in Plymouth in 1884 and resided there until 1887.

From 1887 until his death in 1906, Weldon's work was centered around the development of a fuller understanding of marine biological phenomena and, in particular, the examination of the relationship between various organs of crabs and shrimps, to determine selective death rates in relation to the laws of growth. During his first five years at the Marine Biological Station, Weldon's investigations were directed to the study of classification, morphology, and the development of decapod crustacea. His only work on invertebrate morphology contained an account of the early stage of segmentation and the building of the layers of shrimp. Weldon was both a master of histological techniques and a powerful and accurate draftsman. In 1889 he succeeded E. Ray Lankester in the Joddrell Chair of Zoology at University College London.

During this time Weldon read Francis Galton's *Natural Inheritance*. In this book Galton* had shown that the frequency distributions of the average size of certain organs in man, plants, and moths were normally distributed. Similar investigations had been pursued by the Belgian statistician, Adolphe Quetelet*, whose work was confined to "civilized man." Weldon was interested in investigating those variations in organs in a species living in a wild state, acted upon by natural selection and other destructive influences.

Writing on heredity in 1889, Galton had predicted that selection would not change the shape of the normal distribution*; he expected that his frequency distributions would remain normally distributed in all cases, whether or not animals were under the action of natural selection. Around this time, Weldon began to study the variation of four organs in the common shrimp (*Crangon vulgaris*), and he collected five samples from waters fairly distant from Plymouth. His statistical analysis, published in 1890, confirmed Galton's prediction. Shortly after the paper was published, Weldon was elected a Fellow of the Royal Society.

During the Easter vacation of 1892, Weldon and his wife collected 23 measurements from 1000 adult female shore crabs (*Carcinus mœnas*) from Malta and the Bay of Naples. Weldon discovered that all but one of the 23 characters he measured in the Naples