BU-688-M

COCHRAN-LIKE AND WELCH-LIKE APPROXIMATE SOLUTIONS TO THE PROBLEM OF COMPARISON OF MEANS FROM TWO OR MORE POPULATIONS WITH UNEQUAL VARIANCES

Barbara A. Grimes, Cornell University

Walter T. Federer, Cornell University

The problem of comparing independent sample means arising from populations with unequal variances has been under consideration for many years for the case of two populations. Historically, this problem has come to be known as the Behrens-Fisher problem. This paper is concerned with the comparison of more than two independent sample means arising from populations with unequal variances. In this extension of the Behrens-Fisher problem, the following notation is used. Let π_i represent a normal population with mean μ_i and variance σ_i^2 for i=1,···,v. From each of the v populations we draw an independent random sample. Let n_i represent the size of the i^{th} sample, and X_{ij} represent the j^{th} observation from the i^{th} sample with j=1,···,n_i and i=1,···,v. X_{ij} is distributed normally with mean μ_i and variance σ_i^2, denoted $X_{ij} \sim N(\mu_i,\sigma_i^2)$. The usual unbiased estimators of μ_i and σ_i^2 are respectively $\bar{X}_i = \Sigma_{j=1}^{n_i} X_{ij}/n_i$ and $s_i^2 = \Sigma_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2/(n_i-1)$ with $n_i-1 = f_i$ degrees of freedom. Denote the tabulated value of Student's t distribution for h degrees of freedom at the α percentage point by $t_\alpha(h)$. The range of summation will be omitted whenever it is clear from the context.

We are concerned with a linear contrast of k of v sample means, say $\Sigma_{i=1}^k c_i\bar{X}_i$ where the c_i's are real numbers such that $\Sigma_{i=1}^k c_i = 0$. The distribution of the sample statistic $d = (\Sigma_{i=1}^k c_i\bar{X}_i)/(\sqrt{\Sigma_{i=1}^k c_i^2 s_i^2/n_i})$ has been considered for the case k=2 by both Welch (1938) and Cochran (1964). For k=2, Cochran (1964) suggested that d could be compared to an approximate critical value t'_α, which is obtained as a weighted sum of Student's t values, namely, $t'_\alpha = (w_1 t_\alpha(f_1) + w_2 t_\alpha(f_2))/(w_1+w_2)$ where $w_i = s_i^2/n_i$. A natural extension for the case $k \geq 2$ is $(\Sigma_{i=1}^k |c_i| w_i t_\alpha(f_i)) \div (\Sigma_{i=1}^k |c_i| w_i) = t'_\alpha$. When $f_i = f$ for all i=1,···,k samples in the contrast, note that the critical value reduces to $t'_\alpha = t_\alpha(f)$.

Welch (1938) proved that the statistic d is distributed approximately as a Student's t, denoted $d \sim t$, for the case k=2 and stated that the extension for $k \geq 2$ is readily available using his method of proof. We have done this in the following theorem; the proof for the general case is detailed in Grimes (1979).

Theorem 1. Let $X_{ij} \sim N(\mu_i,\sigma_i^2)$ where the X_{ij}'s are independent. Let $d = \Sigma c_i\bar{X}_i/\sqrt{\Sigma c_i^2 s_i^2/n_i}$ where $\Sigma c_i = 0$ and $\bar{X}_i$ and s_i^2 are the usual unbiased estimators of μ_i and σ_i^2 respectively. Then under the null hypothesis $\Sigma c_i\mu_i = 0$, d is distributed approximately as a Student's t, with degrees of freedom, b given by

$$b = \left(\sum_{i=1}^k \frac{c_i^2\sigma_i^2}{n_i}\right)^2 \Big/ \left(\sum_{i=1}^k \frac{c_i^4\sigma_i^4}{n_i^2(n_i-1)}\right).$$

Since the σ_i^2 are usually unknown, we must obtain an estimator of b. Two estimators have been studied. The first which will be denoted by b1 is given by replacing σ_i^2 in the expression for b by the sample estimator s_i^2. Letting $\lambda_i = c_i^2/n_i$ and $f_i = n_i-1$, this yields $b1 = (\Sigma\lambda_i s_i^2)^2/(\Sigma\lambda_i^2 s_i^4/f_i)$. The second which will be denoted by b2 is given by $b2 = [(\Sigma\lambda_i s_i^2)^2 - 2(\Sigma\lambda_i^2 s_i^4/(f_i+2))]/\Sigma\lambda_i^2 s_i^4/(f_i+2)$. The following result can be established for b2.

Theorem 2. The numerator of b2 is an unbiased estimator of $(\Sigma\lambda_i\sigma_i^2)^2$ and the denominator is an unbiased estimator of $\Sigma\lambda_i^2\sigma_i^4/f_i$.

The size or significance level of a test is defined to be the probability of the test rejecting a hypothesis given that the hypothesis is actually true. For various situations of sample size and variance imbalance, the tests based on b1, b2, and t'_α were evaluated in terms of their size by computer simulation techniques. Empirical distributions of d consisting of 500 points each were compiled for various sample size, contrast, and variance structure combinations.

The actual significance level was estimated for each of the Welch-like approximations by the following method. Given a nominal significance level, α_N, and a calculated value for degrees of freedom, say $b1_j$, a value a_j was calculated such that $P(|T| > a_j) = \alpha_N$ where T is a variable distributed as Student's t with $b1_j$ degrees of freedom. Letting $R_j = 1$ if $|d_j|$ is greater than a_j and 0 if $|d_j|$ is less than a_j, the estimate of the actual significance level was given by $\hat{\alpha}_A = \Sigma_{j=1}^{500} R_j/500$. This was done for several nominal levels.

Cochran's approximation for the equal sample size case gives a critical value which does not depend upon sample quantities. To assess his approximation for this case a different method was used. The empirical distribution of d was compared, as described in the next paragraph, to t distributions with degrees of freedom ranging from $\min(n_i-1)$ to $\Sigma_{i=1}^k(n_i-1)$. For some of the larger sample sizes the upper range of comparison was extended to infinity. It has been established [Mickey and Brown (1966)] that in the two sample case the distribution of d is bounded by the t-distributions with the degrees of freedom given above, and as the smaller of the two sample sizes approaches infinity, the distribution of d approaches a standard normal distribution.

Given that the number of observations in the empirical distribution is M, an expected cumulative distribution was calculated as follows. For significance level α_N and degrees of freedom equal to h,

the number of observations expected to be greater in absolute value than the corresponding Student's t value, r_h, is equal to α_N times M. This expected number, E_{α_N}, was then compared to the actual number of $|d_j|$'s greater than r_h denoted O_{α_N} for α_N = .5, .4, .2, .1, .05, .025, and .01. Since each empirical distribution was to be compared to a large number of t distributions, a summary statistic was needed to indicate which t distribution gave the "best fit" to each empirical distribution. For each h, the summary statistic, fit, was calculated as follows: fit = $\Sigma_{\alpha_N}(O_{\alpha_N}-E_{\alpha_N})^2/E_{\alpha_N}$ where O_{α_N} and E_{α_N} are as defined above and the sum is taken over the values of α_N given above. The t distribution for which fit had the smallest value was said to give the "best fit" to the empirical distribution under consideration. For the unequal sample sizes case, the actual significance level of the test based on the Cochran-like approximation, t'_α, was estimated as described for the Welch-like estimators.

Since results are available in the literature for the comparison of two sample means from populations with unequal variances, this case was used as a check on the simulation study. Previously established results including the conservativeness of Cochran's approximation except in the case of small sample size combined with a severe imbalance in variance were verified.

The main point of our research, however, was to determine the behavior of Cochran-like and Welch-like statistics when k > 2 means are involved in a contrast. A description of variance imbalance is no longer as simple as in the two means case. The amount of variance imbalance can be described in several ways. First the range of variances of the means under consideration is important. A measure of this type of imbalance is the ratio VIM1 = $\sigma^2_{i,min}/\sigma^2_{i,max}$ where $\sigma^2_{i,min}$ is the minimum variance from variance structure i involved in the contrast and $\sigma^2_{i,max}$ is the maximum variance from variance structure i involved in the contrast. This ratio yields values between zero and one. Values close to one indicate mild imbalance, and values close to zero indicate severe imbalance of variances.

A second possibility is that perhaps the variances of the mixed populations is the factor affecting the distribution of d. For example, if our contrast is of the form $\bar{Y}_1 + \bar{Y}_2 - 2\bar{Y}_3$ where $\bar{Y}_i$ is the mean of the sample of size n drawn from population π_i, i=1,2,3, we say that populations π_1 and π_2 are mixed with variance $(\sigma^2_1 + \sigma^2_2)/2$ and π_3 has variance σ^2_3. The ratio of the quantities $(\sigma^2_1+\sigma^2_2)/2$ and σ^2_3 would be another measure of variance imbalance. Following this rationale VIM2 was constructed as follows. Let $G2 = (\Sigma\sigma^2_i I\{c_i > 0\}) \div (\Sigma I\{c_j > 0\})$, and $L2 = (\Sigma\sigma^2_i I\{c_i < 0\})/(\Sigma I\{c_j < 0\})$ where $I\{A\}$ = 1 if A is true and $I\{A\}$ = 0 if A is false. Define VIM2 = min(G2,L2)/max(G2,L2). Then $0 < VIM2 \leq 1$ for all variance structure and contrast combinations and VIM2 = 1 when $\sigma^2_i = \sigma^2$, $\forall$ i. Values of VIM2 close to one will indicate situations of mild variance imbalance, and values close to zero will indicate severe imbalance.

Table 1 gives the contrast matrix and variance structure matrix for which we shall discuss specific simulation results for the equal sample sizes case. The elements in each row of a contrast matrix are contrast coefficients. The elements on each row of a variance structure matrix are the variances, σ^2_i, of the populations, π_i, from which samples are drawn. Samples of size 5, 10, 20, and 40 were drawn in this study.

Table 2 gives the values of VIM1 and VIM2 for the contrast matrix I and variance structure matrix I combination and the degrees of freedom for the Student's t distribution which, under each variance structure, contrast, and sample size combination, gave the best fit. For a specific variance structure, VIM1 takes on the same value for each contrast since all the means are involved in each contrast. VIM1 does differentiate between the three variance structures, indicating that variance structure 3 is the case of most severe imbalance and variance structure 1 is the case of least severe imbalance. For samples of size 5, 10, and 20, the contrasts in combination with variance structure 1 do have for the most part larger associated degrees of freedom than the contrasts in combination with variance structure 3.

The measure VIM2 takes on a wide range of values. VIM2 takes on its minimum value under the contrast 1 and variance structure 3 combination. For samples of size 5 and 10, this combination does have the smallest associated degrees of freedom as determined by the fit criterion. For samples of size 20 and 40, this effect is no longer seen. Thus in the case of equal sample sizes as sample size increases, the effect of inequality of variances on the distribution of d appears to diminish.

The approximation, t'_α, reduces to a Student's t based on n-1 degrees of freedom in the equal sample sizes case. Comparing n-1 to the degrees of freedom given in Table 2, the approximation appears to be an underestimate, and hence conservative. Only in the cases of sample size equal to 5 and VIM2 = .188 and .375 does n-1 seem a reasonable approximation to the desired degrees of freedom as indicated by the fit criterion.

On the other hand, the degrees of freedom for the equal variances case, kn-k, appears to be a reasonable estimate in most cases. This estimate gives 12 for n = 5, 27 for n = 10, 57 for n = 20, and 117 for n = 40. Comparing these estimates with the values in Table 2, we see that for the majority of variance structure, contrast, and sample size combinations, the estimates are close to those degrees of freedom which gave the best fit.

Ideally, we would like to combine the desirable features of the Cochran-like estimator and the equal variance degrees of freedom. Table 3, which gives the average degrees of freedom and variance given by the Welch-like approximations b1 and b2 under each variance structure and contrast combination for samples of size 10 indicates that b1 and b2 behave in such a manner. The behavior was similar for other sample sizes. Both b1 and b2 fall between the values given by the Cochran-like approximation and the equal variances situation. They each take on their smallest values in the row of

each display associated with the first contrast. VIM2 also takes on its smallest values in this row. Hence the behavior of approximations b1 and b2 appears to be reflecting the type of variance imbalance measured by VIM2.

In Table 4 we compare the nominal significance level, α_N, to the achieved significance levels, $\hat{\alpha}_A$, for each of the Welch-like approximations for variance structures 1 and 3 and samples of size 10 and 40. For samples of size 10, the significance levels achieved by b1 and b2 are close to the nominal levels. b1 is a bit on the conservative side, while $\hat{\alpha}_A$ for b2 is generally larger than the corresponding α_N. b1 and b2 do better for samples of size 40. This is due to the fact that, for large degrees of freedom, the t distributions do not differ greatly in their critical values. It should also be noted that as n increases b1 and b2 yield the same values for $\hat{\alpha}_A$. This can be explained by writing b2 = R(b1) - 2 where $R = (\Sigma\lambda_i^2 s_i^4/(n_i-1))/(\Sigma\lambda_i^2 s_i^4/(n_i+1))$. For large n_i, n_i-1 is approximately the same as n_i+1 so R is approximately equal to 1. Hence, b1 differs from b2 by approximately 2. For large values of degrees of freedom the critical values given by the Student's t distribution differ little over a range of 2 degrees of freedom.

Best fit degrees of freedom and estimated actual significance levels for b1 for the contrast matrix II, variance structure matrix II, and equal sample sizes of 10 combinations are given in Table 5. Each number displayed is the average value for 5 different simulations with the standard deviation of a single observation given in parentheses. The experiment was repeated in this case to give us some idea of the accuracy of our results. Despite the variability in the best fit criteria, we still see a difference between the three contrasts. The variability is not great in the estimated significance levels for b1 so 500 seems to be a reasonable number of points for the empirical distribution. The variability associated with b2 is similar, and so is not presented here. Other results for the four sample case were comparable to the previous results for the three sample case.

Finally, we considered the case when the means being compared were estimated from samples of unequal sizes. Two extreme situations were considered. First, in what we called a natural design, the ratio σ_i^2/n_i was taken to be a constant. Conversely, in what we called an unnatural design, σ_i^2 times n_i was taken to be a constant. By looking at the behavior of the approximations for these two situations we hoped to gain an idea of their behavior in intermediate situations. The experiment involved three populations with variances $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, and $\sigma_3^2 = 6$. For the natural design, D1, we had $n_1 = 5$, $n_2 = 10$, and $n_3 = 30$. For the unnatural design, D2, we had $n_1 = 30$, $n_2 = 15$, and $n_3 = 5$. Contrast matrix I from Table 1 was used in the generation of the empirical distributions of d. Looking at Table 6 we see that, for each contrast, the degrees of freedom indicated by the fit criterion is larger in the natural design case than in the unnatural design case. Both b1 and b2 are close to the values given by the fit criterion except for the natural design and contrast 1 combination.

In Figure 1 the achieved significance levels, $\hat{\alpha}_A$, for the tests based on b1 and t'_α are plotted as the ordinate and the nominal significance levels, α_N, as the abscissa for both designs. We see that for both designs t'_α is quite conservative when compared with b1. However, b1 yields significance levels closer to the nominal level in most cases. b2 was not plotted, since when both b1 and b2 yield $\hat{\alpha}_A > \alpha_N$, $\hat{\alpha}_A$ for b1 is less than $\hat{\alpha}_A$ for b2.

In summary we can list the following results of our study:

1. As expected, equality of sample sizes tempered the effect of unequal variances on the distribution of d.

2. An unexpected result was that as sample size increased the effect of unequal variances was diminished. However, upon reflection, this should be the case because looking at the denominator of d we see that the quantity s_i^2 appears divided by n_i. Hence, for large n_i the inequality of variances will not have as strong an effect on the distribution of d as for small n_i.

3. The Cochran-like approximation, t'_α, was more conservative in all cases than the Welch-like approximations b1 and b2.

4. When choosing between the two Welch-like estimators, b1 is the preferred estimator.

5. The choice between t'_α and b1 is not clear-cut. If one wishes to use a conservative test, these results recommend the use of t'_α. If one is more interested in detecting differences between means and a false detection is not very costly, a test based on b1 would be recommended.

6. The question of how to best measure variance and sample size imbalance remains open.

REFERENCES

Cochran, W. G. Biometrics 20, (1964), 191-195.

Grimes, B. A. Master's Thesis, Cornell University, (1979).

Mickey, M. R. and Brown, M. B. Annals of Math. Stat. 37, (1966), 639-642.

Welch, B. L. Biometrika 29, (1938), 350-362.

Welch, B. L. Biometrika 34, (1947), 28-35.

Table 1. Contrast and variance structure matrices.

Contrast Matrix I

$$\begin{bmatrix} 1 & 1 & -2 \\ 1 & -2 & 1 \\ -2 & 1 & 1 \end{bmatrix}$$

Contrast Matrix II

$$\begin{bmatrix} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{bmatrix}$$

Variance Structure Matrix I

$$\begin{bmatrix} 1.0 & 1.5 & 2.0 \\ 1.0 & 2.0 & 4.0 \\ 1.0 & 4.0 & 8.0 \end{bmatrix}$$

Variance Structure Matrix II

[1.0 2.0 6.0 10.0]

Table 2. Values of VIM1 and VIM2 and "best fit" degrees of freedom for simulation using contrast matrix I and variance structure matrix I.

VIM1

Contrast	Variance Structure 1	2	3
1,2,3	.500	.250	.125

VIM2

Contrast	Variance Structure 1	2	3
1	.625	.375	.188
2	1.000	.800	.444
3	.571	.333	.200

Sample size: 5

Range of Comparison: [4,12]

Contrast	Variance Structure 1	2	3
1	8	6	6
2	9	9	8
3	11	10	11

Sample size: 10

Range of Comparison: [9,27]

Contrast	Variance Structure 1	2	3
1	26	21-22	19
2	24-26	25-27	26-27
3	21	26	26-27

Sample size: 20

Range of Comparison: [19,60]

Contrast	Variance Structure 1	2	3
1	45-40	55	45
2	50	35	45
3	60	60	55-60

Sample size: 40

Range of Comparison: [40,∞]

Contrast	Variance Structure 1	2	3
1	70-80	80	120
2	∞	120	∞
3	60	100	100

Table 3. Sample average and variance (in parentheses) of b1 and b2 under contrast matrix I and variance structure matrix I for samples of size 10.

Contrast	Variance Structure 1	2	3
1	15.7 (11.46)	13.3 (6.68)	12.5 (4.71)
2	17.8 (11.80)	18.1 (8.11)	16.9 (5.93)
3	21.8 (10.42)	22.9 (6.95)	22.5 (9.73)

Contrast	Variance Structure 1	2	3
1	17.1 (17.11)	14.2 (9.97)	13.3 (7.04)
2	19.8 (17.63)	20.2 (12.11)	18.7 (8.86)
3	24.6 (15.56)	26.0 (10.38)	25.5 (14.53)

Table 4. Achieved significance levels for b1 and b2 under contrast matrix I and variance structure matrix 1.

			Nominal Significance Levels							
			.01	.025	.05	.1	.01	.025	.05	.1
			Achieved Significance Levels[1]							
n	Variance Structure	Contrast	Approximation b1				Approximation b2			
10	1	1	.004	.018	.048	.088			.050	
		2	.008	.014	.044	.098		.016	.046	.102
		3	.006	.028	.056	.108		.030		.108
	3	1	.006	.018	.046	.090		.020		.092
		2	.008	.016	.044	.092		.018	.046	.094
		3	.008	.026	.054	.088		.028	.056	.092
40	1	1	.012	.024	.050	.092				
		2	.010	.020	.048	.096				.098
		3	.014	.030	.052	.106				
	3	1	.010	.026	.048	.084				
		2	.008	.018	.044	.086				
		3	.008	.026	.048	.098				

[1] A blank entry for approximation b2 indicates that b1 and b2 achieve the same significance level.

Table 5. Average best fit degrees of freedom and average achieved significance levels for b1 and contrast matrix II, variance structure matrix II and samples of size 10. Standard deviation is given in parentheses.

Contrast		
1	2	3
14.0 (3.1623)	24.2 (7.5961)	28.9 (8.3096)

	Nominal Significance Levels			
Contrast	.01	.025	.05	.10
1	.0108 (.00228)	.0228 (.00228)	.0440 (.00424)	.1028 (.00390)
2	.0084 (.00358)	.0244 (.00518)	.0556 (.00740)	.1044 (.01417)
3	.0064 (.0026)	.0188 (.00657)	.0480 (.00678)	.0984 (.01099)

Table 6. Best fit degrees of freedom and average values of b1 and b2 for the natural (D1) and unnatural (D2) designs.

	Design	
Contrast	D1	D2
1	22-25	5
2	19	11
3	10	8

Sample Average and Variance

Approximation b1

	Design	
Contrast	D1	D2
1	36.24 (29.93)	4.53 (1.05)
2	17.32 (13.42)	9.99 (16.57)
3	11.20 (32.70)	7.68 (15.47)

Approximation b2

	Design	
Contrast	D1	D2
1	40.77 (21.73)	4.78 (1.62)
2	19.89 (22.54)	12.08 (28.82)
3	14.03 (52.61)	9.30 (27.64)

Figure 1. Achieved significance levels versus nominal significance levels for tests based on b1 and t'_α for the contrast matrix I and natural and unnatural design combinations.