MAURI ORA, 1979, 7: 139-142

OBTAINING THE OVERALL MEAN AND VARIANCE

FOR COMBINED SAMPLES (NOTE)

K.W. DUNCAN

Department of Zoology, University of Canterbury,
Christchurch, New Zealand

ABSTRACT

Formulae are given for combining means and variances from independent samples, and the advisability of pooling or combining samples is discussed.

It is sometimes necessary to calculate overall means and variances for a set of samples when only the mean, variance (or standard deviation) and sample size is known for each sample. For example, to obtain a representative mean clutch size and variance for a particular bird species it may be necessary to combine published means and variances where the original data, or even the sums and sums of squares, are not given. Simply calculating the mean of means is not valid for all but the special case where the sample sizes are identical. Likewise, calculating the mean of variances to obtain an overall variance does not give the true value. For the data in Table 1 calculating the mean of means and variances gives the incorrect values of 12.8 and 36.52 respectively.

TABLE 1. TRIAL DATA AND EXAMPLE OF WORKINGS.

	Sample 1		Sample 2
	5		10
	4		15
	3		25
	2		30
	6		23
Sample size (N_1)	5		41
Mean ($\bar{X}_1$)	4.0		15
Variance (s_1^2)	2.0		17
Variance with			19
Bessel's correction			21
applied ($\hat{s}_1^2$)	2.5	Sample size (N_2)	10
		Mean ($\bar{X}_2$)	21.6
		Variance (s_2^2)	71.04
		Variance with Bessel's correction applied ($\hat{s}_2^2$)	78.93

Table 1 continued:

'Mean of means' = (4 + 21.6)/2 = 12.8 incorrect
'Mean of variances' = (2 + 71.4)/2 = 36.52 incorrect

Overall mean $\bar{X} = \dfrac{\Sigma N_i \bar{X}_i}{\Sigma N_i}$

$$= \frac{5(4) + 10\ (21.6)}{5 + 10} = 15.73 \qquad (1)$$

Overall variance $s^2 = \dfrac{\Sigma\ (N_i(s_i^2 + \bar{X}_i^2))}{\Sigma N_i} - \bar{X}^2$

$$= \frac{5(2 + 4^2) + 10\ (71.04 + 21.6^2)}{5 + 10} - 15.73^2 \qquad (2)$$

$$= 116.97$$

Applying Bessel's correction: $s^2 = Ns^2/\ (N-1)$

$$= 15\ (116.97)/14$$

$$= 125.33$$

A note of caution should be sounded regarding the wisdom of calculating overall means and variances since the overall estimates may not have much biological meaning. The samples may be so different from each other (i.e. heterogeneous) that they should not be combined since the mean and variance of the combined sample would not be representative of any natural population. When in doubt about sample heterogeneity apply a variance ratio test (Sokal and Rohlf 1969: p.186) to the variances; if that is not significant test the means with Student's t. If either or both of these two tests indicate that the samples come from different populations be careful about combining the samples.

Where the samples do come from the same population, or if there are over-riding biological reasons for obtaining combined estimates of the mean and variance, then the procedures below may be followed to obtain 'overall' means and variances for the pooled samples.

OVERALL MEAN

The correct approach is to weight each sample mean ($\bar{X}_i$) by the sample size (N_i). Since

$\bar{X}_1 = A_1/N_1$ and $\bar{X}_2 = A_2/N_2$

where A_i is the sum of the ith sample data values (i.e. ΣX).

Then $A_1 = N_1\bar{X}_1$ and $A_2 = N_2\bar{X}2$.

Thus the grand mean, $\bar{X}$, is given by

$$\bar{X} = \frac{A_1 + A_2}{N} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2}{N_1 + N_2} .$$

Example computations are given in Table 1. Note that the procedure readily extends to three (or more) samples.

For the three sample case:

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3}{N_1 + N_2 + N_3}$$

The general formula, therefore is:

$$\bar{X} = \frac{\Sigma(N_i\bar{X}_i)}{\Sigma N_i} \quad .$$

OVERALL VARIANCE

This is slightly more complicated. It is necessary first to obtain the sum of squares, T, using the sample variance, s^2, the sample mean, $\bar{X}$ and sample size, N.

Since $s_i^2 = T_i/N_i - \bar{X}_i{}^2$

the sum of squares is given by $T_i = N_i(s_i{}^2 + \bar{X}_i{}^2)$.

Calculate the sum of squares for each sample. For the data in Table 1 we have,

$$T_1 = 5\ (2 + 4^2) = 90 \quad ,$$

and $T_2 = 10\ (71.04 + 21.6^2) = 5376$.

Add these to get the total sum of squares,

$$\Sigma X^2 = T_1 + T_2 = 90 + 5376 = 5466.$$

Substitute this in the variance equation:

$$\begin{aligned} s^2 &= \Sigma X^2/N - \bar{X}^2 \\ &= 5466/(5 + 10) - 15.73^2 \\ &= 116.97 \end{aligned}$$

(Note that the mean, $\bar{X}$, used here is the overall mean calculated using equation 1 of Table 1.)

Apply Bessel's correction to obtain $\hat{s}^2$ if desired:

$$\hat{s}^2 = s(n/(n-1)) = 116.97\ (15/14) = 125.33 \quad .$$

There are no set rules as to whether or not Bessel's correction should be applied, but, as a general guide, if the aim in calculating a sample variance, whether for a single sample or a group of samples, is to indicate the population variance then Bessel's correction should be applied; $\hat{s}^2$ is a better estimate of the population σ^2 than is s^2. Note, however, that the larger the sample size the smaller the difference between $\hat{s}^2$ and s^2, so Bessel's correction need not be applied to large samples.

The formula for the overall variance given in Table 1 is quite general and can be applied to two, three or more samples.

When using published variances with small (< 200) samples, be careful to check whether or not Bessel's correction has been applied. Most calculators which calculate the variance or standard deviation automatically do apply Bessel's correction (some, such as the Hewlitt Packard 45, do it incorrectly). For such cases, undo the correction by multiplying the published variance for the ith sample by $(N_i - 1)/N_i$ before using the values in Equation 2 of Table 1.

Snedecor and Cochran (1967: pp.104-106) give a solution for the two sample pooled variance case which yields slightly incorrect values for the overall variance. Their pooled variance formula is correct for use in testing with Student's t, but it is not intended for obtaining 'overall' variances. If their formula is used for this purpose on the data in Table 1 the slightly incorrect value of 126.12 is obtained.

ACKNOWLEDGEMENTS

I would like to thank Mr Peter Smith for drawing my attention to this problem and Dr Larry Field for his valued comments on the manuscript.

LITERATURE CITED

SNEDECOR, G.W. and COCHRAN, W.G. 1967. *Statistical Methods* (6th Edition). Iowa State University Press, Ames, Iowa. 593 pp.

SOKAL, R.P. and ROHLF, F.J. 1969. *Biometry*. W.H. Freeman, San Francisco. 776 pp.