ABACUS

VOL. 26 1998

Combined Variance of Two Independence Samples

Isaac Olayiwola Oshungade,
Department of Statistics,
University of Ilorin,
Ilorin - Nigeria.

Summary

This article discusses four methods of calculating the combined variance for two independent samples if the assumptions are not clearly stated.

1. INTRODUCTION

We consider the problem of how to combine several sample means and variances to obtain a single unbiased estimates of the population mean and variance. Suppose there are two independent samples made up as follows, the first sample has a mean of 3 and variance of 2 and the second sample has a mean of 7 and variance of 5. We can be asked to find the mean and variance of the combined samples.

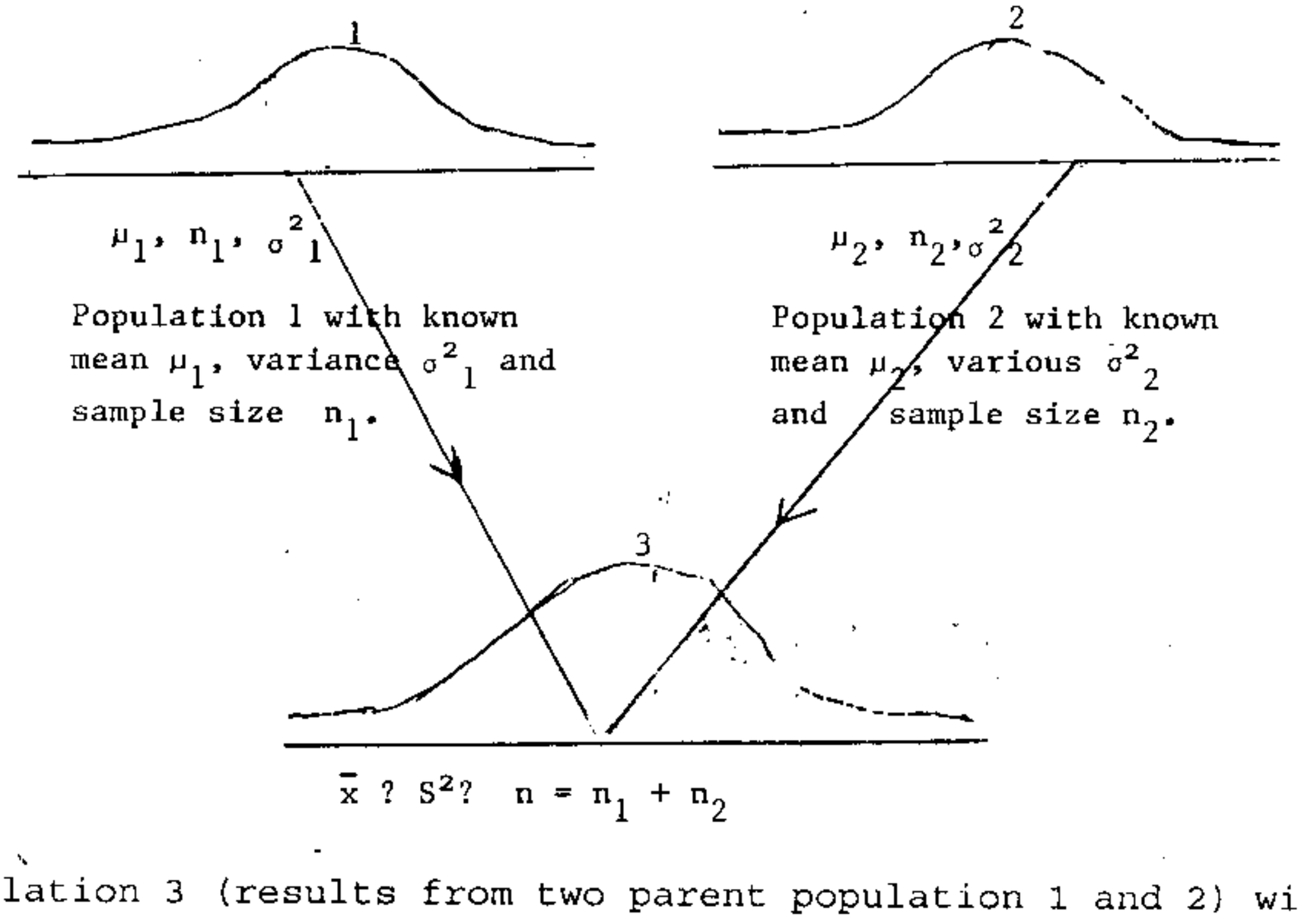
There are occassions when this is definitely desirable. Hoel (1971) has indicated that such a problem would arise in Quality Control work if one wished to obtain an unbaised estimate of the average production and variability of a manufacturing process is measured by σ^2 and had available a number of daily estimate of the variability.

The calculation of the mean of a combined samples creates no problem as it can be easily calculated as

$$\bar{x} = \frac{\sum n_i \bar{x}_i}{\sum n_i}$$

where n_i is the sample size for the ith sample and $\bar{x}_i$ is the sample mean for the ith sample. However to calculate the variance of the combined samples there are about four approaches of doing this. The problem is: Which of these four methods is most appropriate. We also asked ourselves whether is any significant difference between the estimated variances of each of the four approaches. In addition, we could ask which of the approaches has a close or unbiased estimate and under what conditions each of them should be adopted.

The problem above can be illustrated with a simple diagram as shown below.



Population 3 (results from two parent population 1 and 2) with unknown mean $\bar{x}$ and variance S^2 but with a known sample size $n = n_1 + n_2$.

The four approaches or methods are discussed and illustrated in sections two and three of this article.

2. Methods of Estimating the Variance of a Combined Samples.

We can estimate the variance of a combined samples by using any of the following methods viz

(a) **Calculating from the basic formula with the set of values involved**. That is

$$\sigma^2_{1n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

where $n = n_1 + n_2$

(b) **Method of Pooled variance with some degree of fredom**

$$\sigma^2_{2n} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$where \quad S_i^2 = \frac{1}{n_{i-1}}\Sigma(x_i - \bar{x}_i)^2 \qquad i = 1,2$$

(c) **Direct sum of variance of Independent Samples.**

$$\sigma^2_{3n} = \sigma^2_{n_1} + \sigma^2_{n_2}$$

From statistical theory the variance of two Independent Samples A and B will be Variance (A+B) = Variance A + Variance B that is the variance of a sum of Independent variables equals the sum of their variances.

(d) Combined Variance taking care of the deviation from the combined mean $\bar{x}$, that is

$$\sigma^2_{4n} = \frac{n_1(\sigma_1^2 + C_1^2) + n_2(\sigma_2^2 + C_2^2)}{n}$$

Where $C_i = |\bar{x} - x_i|$ and $\bar{x}$, $\bar{x}_i$ are the means for the combined and different ith samples respectively.

The four methods have the following assumptions. For σ^2_{1n} we could assume $\bar{x}_1 = \bar{x}_2$ and $\sigma^2_1 = \sigma^2_2$. For σ^2_{2n} it is assumed $\sigma^2_1 = \sigma^2_2$ that is the two samples belong to the same population and the two variances are not significantly different. The third method σ^2_{3n} assumed that the two samples are strictly independent and may or may not come from the same population. For σ^2_{4n} we assume that the variances are different that is $\sigma^2_1 \neq \sigma^2_2 \neq \sigma^2_k$. So also their means $\bar{x}_1 \neq \bar{x}_2 \neq \bar{x}_k$.

The four methods σ^2_{1n}, σ^2_{2n}, σ^2_{2n}, σ^2_{3n} and σ^2_{4n} would be examined in two ways. First we need to examine the unbiasedness of each of them and secondly their easiness or difficulty of calculation in real life.

$E(\sigma^2_{1n}) = \frac{n-1}{n}\sigma^2$ hence σ^2_{1n} is heavily biased and this bias can be corrected by multiplying it by the factor n/n-1 σ^2_{4n} is also heavily biased.

σ^2_{2n} is unbiased as $E(\sigma^2_{2n}) = \sigma^2$

σ^2_{3n} is unbiased if independent samples are involved. We have assumed that there is no correlation between the two samples. Positive correlation increased the variance of the sum, negative correlation decreases it

The easiest of all to calculate or compute is σ^2_{3n} and the most diffcult or more time consuming from the given information is σ^2_{1n}. However, σ^2_{4n} involves more calculation than in σ^2_{2n}. The imposition of n_1+n_{2-2} degree of freedom on σ^2_{2n} implies that the two sample means are held constant or fixed and has $n_1+n_2 - 2$ independent variates instead of $n = n_1 + n_2$. Thus each estimate variance is weighted by the number of degree of freedom available for its calculation. The third method σ^2_{3n} is unaffected by changes in size of sample.

From the different calculations, σ^2_{4n} gives the same result as σ^2_{1n}. In fact, σ^2_{4n} is an expansion of σ^2_{1n}.

The formulae of σ^2_{2n} and σ^2_{4n} are for cases of unequal sample sizes. For equally weighted samples, that is, cases with $n_i = n$ for i = 1,2.....k the formulae are modified. For example the combined mean

$$\overline{X} = \frac{\overline{X}_1+X_2+\ldots\ldots\ldots+\overline{X}_k}{k}$$

$$\sigma^2_{2n} = \frac{S_1^2 + S_2^2 +\ldots\ldots\ldots+S_k^2}{k}$$

and for

$$\sigma^2_{4n} = \frac{1}{k}\left[\sum_{i=1}^{k}\sigma_i^2 + \sum_{i=1}^{k}C_i^2\right]$$

3. Illustrations of the Four Methods

We consider the problem stated in our introduction which is reproduced below

Sample	Mean	Variance	Sample Size
A	3	3	4
B	7	5	6

As noted in the introduction the combined mean

$$\bar{x} = \frac{4 \times 3 + 7 \times 6}{10} = 5.4$$

The variance for the combined samples using:

$$(a)\ \sigma_{1n}^2 = \frac{\Sigma x^2}{n} - x^2$$

$$\Sigma x^2 = \Sigma x_1^2 + \Sigma x_2^2$$

$$\sigma_1^2 = \frac{\Sigma x_1^2}{n_1} - x_1^2$$

$$2 = \frac{\Sigma_1^2}{4} - 9$$

$$\Sigma x_1^2 = 44$$

$$\sigma_2^2 = \frac{\Sigma x_2^2}{n_2} - x_2^2$$

$$5 = \frac{\Sigma x_2^2}{6} - 49$$

$$\Sigma x_2^2 = 324$$

$$\therefore \Sigma x^2 = 44 + 324 = 368$$

$$\sigma_{1n}^2 = \frac{368}{10} - (5.4)^2$$

$$= 7.64$$

$$\sigma_{1n} = 2.7641$$

$$(b)\ \sigma^2_{2n} = \frac{(n_1-1)\,S_1^2 + (n_2-1)\,S_2^2}{n_1 + n_2 - 2}$$

$$with\ S_i^2 = \frac{n_i\sigma_i^2}{n_{i-1}}$$

$$\sigma^2_{2n} = \frac{3 \times \frac{8}{3} + 5 \times 6}{8} = \frac{38}{8}$$

$$= 4.75$$

$$\sigma_{2n} = 2.1794$$

$$(C)\ \sigma^2_{3n} = \sigma_1^2 + \sigma_2^2$$

$$= 2 + 5$$

$$= 7$$

$$\sigma_{3n} = 2.6457$$

$$(d)\ \sigma^2_{4n} = \frac{n_1(\sigma_1^2 + C_1^2) + n_2(\sigma_2^2 + C_2^2)}{n}$$

$$C_1 = 5.4 - 3 = 2.4$$

$$C_2 = 5.4 - 7 = -1.6$$

$$\sigma^2_{4n} = \frac{4(2 + 5.76) + 6(5 + 2.56)}{10}$$

$$\frac{76.4}{10}$$

$$= 7.64 \text{ as in (a) above}$$

$$\sigma_{4n} = 2.7641$$

4. Any Difference in Interpretation?

Does the variance estimated by the different methods differ significantly from each other. That is

$$H_o : \sigma^2_{1n} = \sigma^2_{2n} = \sigma^2_{3n} = \sigma^2_{4n}$$

H_A : Any two or more variances are different.

To test this we use the Bartlett's test when all samples are of the same size n

$$\chi^2 = 2.3026(n-1)(k\log S^2 - \Sigma\log S^2_i)$$

Where k is the number of samples whose variances are being compared. This test is reasonably accurate for
n- 1 >, 5 (see Fraser (1958))

In this case we are comparing four methods i.e k = 4, n is the size of the sample = $n_1 + n_2 = 10$

$\chi^2 = 2.3026\ (9)\ (4\log S^2 - \Sigma \log S^2_i)$.

In this example, from the different methods

$S^2_{in} = 7.64,\ 4.75,\ 7.00$ and 7.64

$\bar{S}^2 = \Sigma S^2_{in}/k = 27.03/4$

$= 6.7575$

$\log S^2 = 0.829786$

$\Sigma \log S^2_i = 3.2880$

$\chi^2 = 2.3026 \times 9(4 \times 0.8298 - 3.2880)$

$= 0.6466$

$\chi^2_{3,4}$ from table $\alpha = 0.05 = 7.815$. Hence

H_0 holds that is, the variances estimated by the different methods are not significantly different from each other. Each of $\sigma^2_{in} i = 1,2,3,4$ is an estimate of the same σ^2. We can therefore conclude that all the variances are homogeneous.

Conclusions

We should note that the estimation of combined or pooled variance always leads to the most advantageous use of all the information provided by the sample data.

It is necessary to emphasise that if the assumptions are not stated we run into a problem of the choice of appropriate and unbiased combined variance.

In terms of biasedness, an examination of the various estimators σ^2_{1n}, σ^2_{2n}, σ^2_{3n} and σ^2_{4n} has shown that σ^2_{1n} and σ^2_{4n} are heavily biased, σ^2_{3n} is less biased and $\sigma^2_{2n}$is unbiased. To remove the biases in σ^2_{1n} and σ^2_{4n} it is necessary to multiply them by n/n-1.

From our example, it could be observed that whichever of the methods are adopted the estimates are close and it has been shown that there is no significant difference among them. It depends on whether we want to estimate the variance or use it purely for a test of hypothesis.

Averaging of variances without considering the numbers of degrees of freedom involved is incorrect. σ^2_{2n} is the appropriate method to use when the population variance is not known and the two samples or more are small sized samples.

In tha analysis of variance, σ^2_{3n} is very useful and is of great importance in assigning errors to various causes.

For a direct total estimates of the variance, for the combined samples, any of σ^2_{1n} or σ^2_{4n} is most appropriate. However in the test of hypothesis σ^2_{2n} is more appropriate.

Care must be taken when combining samples to estimate the true variance when an unbiased estimate is needed. In sampling, the estimation of a variance in a stratified sampling is a kind of combined variances from the independent strata which always give an unbiased estimate of the true variance.

References

1. Hoel, P.G. (1971). 'Introduction to Mathematical Statistics': John Wiley & Sons, New York.

2. Fraser, D.A.A.(1958) 'Statistics - An Introduction' John Wiley & Sons, New York