THE EXPECTATIONS OF MEAN SQUARES

by

R. E. Comstock

Institute of Statistics
Mimeograph Series No. 76
For Limited Distribution

Chapter VI

THE EXPECTATIONS OF MEAN SQUARES

The Expectation of a Variable

If individuals are drawn randomly from a population their average value in terms of any specified measurement will be equal in the long run to the mean for the measurement in the population. We say that the value to be expected on the average is that of the population mean. In fact, in Statistics the expectation of a variable quantity is defined as the mean for such quantities in the population to which the particular variate belongs. For example, let X_1, X_2, ... X_i ..... symbolize the values of the individuals in any univariate population. Then the X's constitute a population of quantities of which the expectation of any one chosen at random is μ_x where μ_x is the population mean. This is stated symbolically as follows:

$$E(X_i) = \mu_x$$

where X_i can be any of the X's depending on the value given i and $E(X_i)$ is read "the expectation of X_i".

As a second example, recall that the population variance is defined as

$$\sigma^2 = \sum_i (X_i - \mu_x)^2/N$$

where σ^2 symbolizes the population variance,

X_i symbolizes the value of any individual quantity in the population,

N is the number of individuals in the population, and

μ_x as before is the population mean.

Thus the variance, σ^2, is defined as the mean of all values, i.e. the population mean, of $(X_i - \mu_x)^2$. In accord with the definition of expectation we see that

$$E(X_i - \mu_x)^2 = \sigma^2 \qquad (2a)$$

or if we wish to represent the deviation of X_i from its population mean by a single symbol, say x_i, we can write

$$x_i = X_i - \mu_x$$

$$E(x_i^2) = \sigma^2 \qquad (2b)$$

As a final example recall that the population covariance of two variables, say X and Y is defined as

$$\sigma_{xy} = \sum_i (X_i - \mu_x)(Y_i - \mu_y)/N$$

where σ_{xy} is the covariance and other symbols have meanings in conformity with those listed above when considering the variance of X. We see that the covariance, σ_{xy}, is defined as the population mean of $(X_i - \mu_x)(Y_i - \mu_y)$ and therefore that

$$E(X_i - \mu_x)(Y_i - \mu_y) = \sigma_{xy} \tag{3a}$$

Again if we set

$$x_i = X_i - \mu_x$$

$$\text{and } y_i = Y_i - \mu_y$$

we can write

$$E(x_i y_i) = \sigma_{xy} \tag{3b}$$

Interest in expectations centers around the fact that by setting observed quantities equal to their expectations we find a basis for unbiased estimation of parameters involved in the expectation. For example, it can be shown that

$$E(X_i - \bar{X})^2 = \frac{n-1}{n}\sigma^2$$

where $\bar{X}$ is the mean of a sample of X's, and

n is the number of individuals in the sample.

It follows that

$$E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n(n-1)}{n}\sigma^2 = (n-1)\sigma^2$$

or

$$E\left[s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] = \sigma^2$$

From this we see that sample variance obtained by dividing the sum of squares by degrees of freedom has σ^2 as its expectation, i.e. that it provides on unbiased estimate of σ^2.

Expectation of a Constant

This is specifically mentioned for completeness. Since a constant, by definition, is a quantity that always has the same value, the expectation of a constant could hardly be anything but that particular value. For example, a population mean is a constant and its expectation is the mean itself. Symbolically, if c is any constant,

$$E(c) = c \tag{4}$$

Expectation of the Product of a Constant and a Variable

Consider the product

$$Y = c\,X$$

where X is a variable and c is a constant. We know that the population mean of Y is $c\,\mu_x$ and, therefore,

$$E(Y = c\,X) = c\,\mu_x = c\,E(X) \qquad (5)$$

In general, the expectation of such a product is the product of the constant and the expectation of the variable.

The Expectation of a Linear Function

Consider the linear function

$$F = a + b + c\,X_1 + X_2$$

in which a, b and c are constants and X_1 and X_2 are variable quantities drawn randomly (but not necessarily independently) from two populations (one, a population of quantities symbolized as X_1, the other a population of quantities symbolized as X_2). Two points are worth special attention.

(1) The specific manner in which F is defined may have the result that values of X_1 and X_2 contributing to different values of the quantity, F, are correlated or on the other hand are independent, i.e. uncorrelated. For example, suppose F is designed to reflect in some special way the height of married couples. Then any single value of F would involve the height of the husband (X_1) and that of his wife (X_2). If the couples are chosen randomly both X_1 and X_2 are random values from their respective populations, but are not necessarily independent in magnitude from one couple to another. In fact, evidence indicates that there is a degree of correlation in stature of man and wife.

On the other hand, suppose F were defined as the height of plants, X_1 as the effect of genotype, and X_2 as the effect of environment on height; and it were known that in the population of plants involved genotypes were distributed randomly with respect to environment. The magnitudes of X_1 and X_2 would vary independently from plant to plant and, therefore, from one value of F to another.

(2) The different variables may actually belong to the same population though it may be useful to think of them as coming from different ones. For example, in the function given above X_1 and X_2 could be a pair of values drawn randomly from the same population, X_1, being the first and X_2 the second drawn of any pair. In this case X_1 and X_2 would vary independently, i.e. be uncorrelated.

Corresponding to every possible pair of values of X_1 and X_2 there is obviously a value of F. These values comprise a population of F's. We know that the mean value of F in that population is

$$a + b + c\ \mu_1 + \mu_2$$

where μ_1 and μ_2 are the population means for X_1 and X_2, respectively. Hence

$$E(F) = \mu_F = a + b + c\ \mu_1 + \mu_2$$

where μ_F is the population mean of F. This serves to demonstrate the general fact that the expectation of a variable quantity that is a linear function of other variables is the same linear function of the expectations of those variables. By this rule

$$E(F) = E(a) + E(b) + E(c\ X_1) + E(X_2)$$

and since

$$E(a) = a$$
$$E(b) = b$$
$$E(c\ X_1) = c\ \mu_1$$
$$E(X_2) = \mu_2$$

We have by substitution

$$E(F) = a + b + c\ \mu_1 + \mu_2$$

as given above.

Expectations of Mean Squares

Any mean square can be written as a linear function in which the variable quantities are the squares of variables, products of a variable with a constant, or products of variables. Hence, the expectations can always be written in terms of what is presented above. This fact will be clarified by examples.

Example 1

Consider the case represented by the analysis of variance for comparing groups of equal size. The form of the analysis is as follows:

Variance Source	d.f.	m.s.
Groups	m-1	M_1
Within groups	m(n-1)	M_2
Total	m n-1	

where m is the number of groups and n is number of individuals within groups. The model on which the analysis is based can be stated symbolically as follows:

$$Y_{ij} = \mu + g_i + e_{ij}$$

where μ is the population mean taken over all groups,

g_i is the effect of the i-th group (the amount by which the population mean for the i-th group deviates from μ), and

e_{ij} is a random effect contributing to the value of Y for the j-th individual in the i-th group (the amount by which the individual deviates from the mean for its group).

One of two assumptions is usually made concerning the groups: (a) that they are random members of a population of groups, or (b) that the ones on which data are taken are of special interest in themselves rather than as a sample from a population. In case (a) the assumption is frequently stated by saying that g_i is considered a random variable, in contrast to case (b) where it is alternatively said that the g_i are considered constant or fixed.

g assumed to be a random variable

We will consider first the case where g_i is considered a random variable. Let

G_i be the sum of Y's for the N individuals of the i-th group, and

T be the sum of Y's for all nm individuals on which data were collected.

Then the mean square for groups is computed as,

$$M_1 = \left[\frac{G_1^2}{n} + \frac{G_2^2}{n} + \cdots \frac{G_m^2}{n} - \frac{T^2}{nm} \right] / m-1$$

This may be considered the product of a constant and a variable where $\frac{1}{m-1}$ is the constant and the quantity in brackets is the variable. Hence, its expectation may be written,

$$E(M_1) = \frac{1}{m-1} E\left[\frac{1}{n}(G_1^2 + G_2^2 + \cdots G_m^2) - \frac{T^2}{nm}\right]$$

Note that $\frac{1}{n}(G_1^2 + G_2^2 + \cdots G_m^2)$ is what we commonly call the "uncorrected sum of squares", that T^2/nm is what we call the "correction factor", and that the whole quantity in brackets is the "corrected sum of squares".

By the rule that the expectation of a linear function is the same function of expectations of the variables in the function, we can write

$$E(M_1) = \frac{1}{m-1}\left[\frac{1}{n}(EG_1^2 + EG_2^2 + \cdots EG_m^2) - \frac{1}{nm} ET^2\right] \qquad (6)$$

Now the separate expectations in the expression can be considered one by one.

Consider EG_i^2. In terms of our model,

$$EG_i^2 = E\left[\sum_{j=1}^{n} Y_{ij}\right]^2 = E\left[Y_{i1} + Y_{i2} + \cdots Y_{in}\right]^2$$

$$= E(n\mu + ng_i + e_{i1} + e_{i2} + \cdots e_{in})^2$$

Squaring and taking expectations term by term this can be written

$$EG_i^2 = En^2\mu^2 + En^2g_i^2 + E(e_{i1} + e_{i2} + \cdots e_{in})^2$$

$$+ E2n^2\mu g_i + E2n\mu\,(e_{i1} + e_{i2} + \cdots e_{in})$$

$$+ E2ng_i\,(e_{i1} + e_{i2} + \cdots e_{in}) \qquad (7)$$

Before going further note, that both the g's and e's are defined as deviations from a mean and hence, that the population mean of both the g's and e's is zero. Thus, $E(g_i) = 0$, $E(e_{ij}) = 0$, $E(g_i^2) = \sigma_g^2$ and $E(e_{ij}^2) = \sigma_e^2$ where σ_g^2 is the population variance of g's and σ_e^2 is the population variance of e's. It is common to assume that all e's are members of the same population and, therefore, that σ_e^2 is homogeneous over all groups. This assumption will be made for the purpose of our example but it should be understood that special cases may arise where the variance of e varies from group to group. It should also be noted that all g's and e's are assumed to

be random members of their populations. The significance of this is that in the population (the population that would be generated by repeating the experiment in identical fashion an infinity of times) the correlation between (1) any two g's, (2) any e's, or (3) any $\underline{g}$ and any $\underline{e}$ would be zero. If the correlation is zero, so also is the covariance and this means that the expectations of all products of two g's, two e's or a $\underline{g}$ and an $\underline{e}$ are all equal to zero. Symbolically this is stated as follows:

$$E(g_i\, g_{i'}) = 0 \qquad (i \neq i')$$

$$E(e_{ij}\, e_{i'j'}) = 0 \qquad (i \neq i' \text{ if } j = j',\ j \neq j' \text{ if } i = i')$$

$$E(g_{i'}\, e_{ij}) = 0 \qquad (\text{either when } i = i' \text{ or when } i \neq i')$$

Now let us consider the several terms of EG_i^2 one by one.

(a) $E\, n^2\mu^2 = n^2\mu^2$. (since $n^2\mu^2$ is a constant)

(b) $E\, n^2 g_i^2 = n^2\, Eg_i^2$ (since n^2 is a constant)

$= n^2\sigma_g^2$ (since $Eg_i^2 = \sigma_g^2$)

(c) $E(e_{i1} + e_{i2} + \cdots e_{in})^2$

$= E(e_{i1}^2 + e_{i2}^2 + \cdots e_{in}^2 + 2e_{i1}\, e_{i2}$

$+ \cdots 2e_{i1}\, e_{in} + \cdots 2e_{i(n-1)}\, e_{in}$

$= n\sigma_e^2$ (since the expectation of each of the e^2's, n of them, is σ_e^2 and that of each product term is zero)

(d) $E2n^2\mu g_i = 2n^2\mu Eg_i$ (since $2n^2\mu$ is a constant)

$=$ zero (since $Eg_i = 0$)

(e) $E2n\mu\,(e_{i1} + e_{i2} + \cdots e_{in})$

$= 2n\mu E\,(e_{i1} + e_{i2} + \cdots e_{in})$ (since $2n\mu$ is a constant)

$=$ zero (since the expectation of all e's is zero and, therefore, that of the sum of any set of e's is also zero)

(f) $E2ng_i\,(e_{i1} + e_{i2} + \cdots e_{in}) = 2nEg_i\,(e_{i1} + e_{i2} + \cdots e_{in})$ (since 2n is a constant)

$=$ zero (since the expectation of the product of any $\underline{g}$ and $\underline{e}$ is zero)

Substituting in (7) in terms of (a) to (f) we find that,

$$EG_i^2 = n^2\mu^2 + n^2\sigma_g^2 + n\sigma_e^2 \tag{8}$$

Now note that nothing in (8) is specific for the particular group in question (i does not appear as a subscript in the right hand member). The significance is that the expectation of G^2 is the same for all groups, that

$$EG_1^2 = EG_2^2 = \ldots\ EG_m^2$$

In order to evaluate $E(M_1)$ it remains only to obtain ET^2.

$$E\,T^2 = E\,(G_1 + G_2 + \ldots\ G_m)^2$$

Substituting for the G's we obtain

$$E\,T^2 = E\Big[nm\mu + n(g_i + g_2 + \ldots\ g_m) + e_{11} + e_{12} + \ldots\ e_{in} + e_{21} + e_{22} + \ldots + e_{2n} \ldots + e_{m1} + e_{m2} + \ldots\ e_{mn}\Big]^2 \tag{9a}$$

Squaring, taking expectations term by term, and moving constants to the left of the sign for expectation (proper because the expectation of the product of a constant and a variable is equal to the product of the constant and the expectation of the variable) we get

$$E\,T^2 = n^2m^2\mu^2 + n^2\,Eg_1^2 + n^2\,Eg_2^2 + \ldots\ n^2\,Eg_m^2 + Ee_{11}^2 + Ee_{12}^2 + \ldots\ Ee_{mn}^2 + \text{product terms of the types } 2n^2m\mu\,Eg_1,\ 2n^2\,Eg_1g_2,\ 2n\,Eg_1\,e_{11},\text{ or } 2Ee_{11}\,e_{12} \tag{9b}$$

Consider the various terms of this expression

(g) $n^2\,Eg_1^2 = n^2\,Eg_2^2 = \ldots\ n^2\,Eg_m^2 = n^2\,\sigma_g^2$ (since $Eg_i^2 = \sigma_g^2$)

(h) $Ee_{11}^2 = Ee_{12}^2 = \ldots = Ee_{mn}^2 = \sigma_e^2$ (since $Ee_{ij}^2 = \sigma_e^2$)

(i) All product terms are of types shown to have zero expectation in the process of developing $E\,G_i^2$.

Substituting in (9b) in terms of (g) to (i) we obtain

$$E\,T^2 = n^2m^2\mu^2 + n^2m\sigma_g^2 + nm\sigma_e^2 \tag{10}$$

Finally substituting in (6) in terms of (8) and (10) we find

$$E(M_1) = \frac{1}{m-1}\left[\frac{m}{n}(n^2\mu^2 + n^2\sigma_g^2 + n\sigma_e^2) - \frac{1}{nm}(n^2m^2\mu^2 + n^2m\sigma_g^2 + nm\sigma_e^2\right]$$

$$= \mu^2\left[\frac{mn-mn}{m-1}\right] + \sigma_g^2\left[\frac{mn-n}{m-1}\right] + \sigma_e^2\left[\frac{m-1}{m-1}\right] = n\sigma_g^2 + \sigma_e^2 \qquad (11)$$

The within group mean square may be computed as follows:

$$M_2 = \frac{1}{m(n-1)}\sum_{i=1}^{m}(Y_{i1}^2 + Y_{i2}^2 + \cdots Y_{in}^2 - \frac{G_1^2}{n})$$

Remembering (a) that the expectation of the product and of a constant and a variable is the product of the constant and the expectation of the variable and (b) that the expectation of a variable that is a linear function of variables is the same function of the expectations of these later variables, we see that

$$E(M_2) = \frac{1}{m(n-1)}\sum_{i=1}^{m}\left[EY_{i1}^2 + EY_{i2}^2 + \cdots EY_{in}^2 - \frac{1}{n}EG_i^2\right] \qquad (12)$$

Consider the expectation of Y_{ij}^2

$$Y_{ij} = \mu + g_i + e_{ij}$$

Therefore,

$$EY_{ij}^2 = E(\mu + g_i + e_{ij})^2$$

Expanding and taking expectations of individual terms separately we obtain,

$$EY_{ij}^2 = E\mu^2 + Eg_i^2 + Ee_{ij}^2 + E\,2\mu g_i + E\,2\mu e_{ij} + E\,2g_ie_{ij} \qquad (13)$$

Taking the terms of this expression separately,

(j) $E\mu^2 = \mu^2$ (because μ^2 is a constant),

(k) $Eg_i^2 = \sigma_g^2$ (by definition when the g's are assumed random),

(l) $Ee_{ij}^2 = \sigma_e^2$ (by definition),

(m) $E\,2\mu g_i = 2\mu\,Eg_i$ = zero (since 2μ is a constant and $Eg_i = 0$),

(n) $E\,2\mu e_{ij} = 2\mu\,Ee_{ij}$ = zero (since 2μ is a constant and $Ee_{ij} = 0$),

(o) $E\,2g_ie_{ij} = 2\,Eg_ie_{ij}$ = zero (since 2 is a constant and $Eg_ie_{ij} = 0$).

Substituting in (13) in terms of (j) to (0) we obtain,

$$EY_{ij}^2 = \mu^2 + \sigma_g^2 + \sigma_e^2 \qquad (14)$$

We have already shown (8) that the expectation of G_i^2 is,

$$E\ G_i^2 = n^2\mu^2 + n^2\sigma_g^2 + n\sigma_e^2 \tag{8}$$

Note that both (14) and (8) are the same for all Y's and G's, respectively (all terms in right hand members are constants). Recognizing this and substituting in (12) in terms of (8) and (14) we obtain,

$$E(M_2) = \frac{m}{m(n-1)}\left[n(\mu^2 + \sigma_g^2 + \sigma_e^2) - \frac{1}{n}(n^2\mu^2 + n^2\sigma_g^2 + n\sigma_e^2)\right]$$

$$= \mu^2\left[\frac{m(n-n)}{m(n-1)}\right] + \sigma_g^2\left[\frac{m(n-n)}{m(n-1)}\right] + \sigma_e^2\left[\frac{m(n-1)}{m(n-1)}\right] = \sigma_e^2 \tag{15}$$

Using (11) and (15) the analysis of variance can now be presented giving the expectations of the mean squares.

Variance Source	d.f.	Expectation of m.s.
Groups	m-1	$\sigma_e^2 + n\sigma_g^2$
Within groups	m(n-1)	σ_e^2
Total	mn-1	

g's assumed to be constants

Differences occasioned by assuming the g's constant rather than random are listed below.

g's random	g's constant
$E\ g_i = 0$	$E\ g_i = g_i$
$E\ g_i^2 = \sigma_g^2$	$E\ g_i^2 = g_i^2$
$Ecg_i = 0$	$Ecg_i = cg_i$

where c is any constant

Other expectations involved in (7), (9b), and (13) are not affected. With the above differences in mind we see that in this case (7) does not reduce to (8) but to

$$E\ G_i^2 = n^2\mu^2 + n^2g_i^2 + 2n^2\ \mu g_i + n\sigma_e^2 \tag{16}$$

In like manner (9b) reduces to

$$E\ T^2 = n^2m^2\mu^2 + nm\sigma_e^2 \tag{17}$$

rather than to (10). The reason why no terms involving g's or the squares or products of g's occurs in (17) is clarified by reference to (9a). Note that the g's enter (9a) in a term that is the sum of the g's for the m groups. In the case where the g's are assumed constant μ is taken as the population mean for the m groups in question. Then, since the g's are defined as deviations from this mean, their sum must be zero. Hence, the term $n(g_1 + g_2 + \ldots. g_n)$ disappears from (9a) and correspondingly terms involving g's disappear from (9b). Finally (13) reduces to

$$E\, Y_{ij}^2 = \mu^2 + g_i^2 + 2\mu g_i + \sigma_e^2 \tag{18}$$

rather than to (14).

Substituting in (6) in terms of (16) and (17) rather than in terms of (8) and (10) we obtain,

$$E(M_1) = \frac{1}{m-1}\left[\frac{1}{n}(mn^2\mu^2 + n^2 \sum_{i=1}^{m} g_i^2 + 2n^2\mu \sum_{i=1}^{m} g_i + nm\sigma_e^2) - \frac{1}{nm}(n^2m^2\mu^2 + nm\sigma_e^2)\right]$$

Keeping in mind that $\sum_{i=1}^{m} g_i = 0$ as pointed out above this reduces to

$$E(M_1) = \mu^2\left[\frac{mn-mn}{m-1}\right] + \frac{n}{m-1}\sum_{i=1}^{m} g_i^2 + \sigma_e^2\left[\frac{m-1}{m-1}\right]$$

$$= \frac{n}{m-1}\sum_{i=1}^{m} g_i^2 + \sigma_e^2 \tag{19}$$

Substituting in (12) in terms of (16) and (18) rather than in terms of (8) and (14) we obtain,

$$E(M_2) = \frac{1}{m(n-1)}\left[(mn\mu^2 + n\sum_{i=1}^{m} g_i^2 + 2\mu n\sum_{i=1}^{m} g_i + mn\sigma_e^2) - \frac{1}{n}(mn^2\mu^2 + n^2\sum_{i=1}^{m} g_i^2 + 2n^2\mu\sum_{i=1}^{m} g_i + mn\sigma_e^2)\right]$$

$$= \mu^2\left[\frac{mn-mn}{m(n-1)}\right] + \sum_{i=1}^{m} g_i^2\left[\frac{n-n}{m(n-1)}\right] + \sigma_e^2\left[\frac{mn-m}{m(n-1)}\right]$$

$$= \sigma_e^2 \tag{20}$$

We have again used the fact that $\sum_{i=1}^{m} g_i = 0$. The analysis of variance with expectations is now as follows:

Variance Source	d.f.	Expectation of m.s.
Groups	m-1	$\sigma_e^2 + \frac{n}{m-1} \sum_{i=1}^{m} g_i^2$
Within groups	m(n-1)	σ_e^2
Total	mn-1	

Example 2

As a variation of example 1 consider the analysis of variance for comparison of groups of unequal size. Let $n_1, n_2 \ldots\ldots n_m$ symbolize the number per group in groups 1, 2, ... m, respectively. The form of the analysis is as follows:

Variance Source	d.f.	m.s.
Groups	m-1	M_1
Within groups	$\sum_{i=1}^{m} (n_i-1)$	M_2
Total	N-1	

where N is the total number of individuals in all groups.

Except for variation in group size the model will be the same as in example 1. We will consider only the case where $\underline{g}$ is considered a random variable. The mean square for groups is computed as

$$M_1 = \frac{1}{m-1}\left[\frac{G_1^2}{n_1} + \frac{G_2^2}{n_2} + \ldots\ldots \frac{G_m^2}{n_m} - T^2/N\right] \tag{21}$$

Referring to (7) and (8) it is clear that

$$E\, G_i^2 = n_i^2 \mu^2 + n_i^2 \sigma_g^2 + n_i \sigma_e^2$$

and hence that

$$E(G_i^2/n_i) = n_i \mu^2 + n_i \sigma_g^2 + \sigma_e^2 \tag{22}$$

T is now equal to

$$N\mu + n_1 g_1 + n_2 g_2 + \cdots n_m g_m + e_{11} + e_{12} + \cdots e_{1n_1}$$

$$+ e_{21} + e_{22} + \cdots e_{2n_2} + \cdots e_{m1} + e_{m2} + \cdots e_{mn_m}$$

Squaring and taking expectations but ommitting terms with expectation zero we obtain,

$$E\,T^2 = N^2\mu^2 + E\,n_1^2 g_1^2 + E\,n_2^2 g_2^2 + \cdots + E\,n_m^2 g_m^2$$

$$+ E\,e_{11}^2 + E\,e_{12}^2 + \cdots + E\,e_{1n_1}^2 +$$

$$\cdots E\,e_{m1}^2 + E\,e_{m2}^2 + \cdots + E\,e_{mn_m}^2$$

Evaluating the separate terms, this becomes

$$E\,T^2 = N^2\mu^2 + n_1^2\sigma_g^2 + n_2^2\sigma_g^2 + \cdots n_m^2\sigma_g^2 + N\sigma_e^2$$

and hence,

$$E(T^2/N) = N\mu^2 + \sigma_g^2 \sum_{i=1}^{m} n_i^2/N + \sigma_e^2 \qquad (23)$$

We have $N\sigma_e^2$ because there are a total of N terms of the type Ee_{11}^2 that are equal to σ_e^2. Writing $E(M_1)$ in terms of (21), (22), and (23) we get

$$E(M_1) = \frac{1}{m-1}\left[\mu^2 \sum_{i=1}^{m} n_i + \sigma_g^2 \sum_{i=1}^{m} n_i + m\sigma_e^2 - N\mu^2 - \sigma_g^2 \sum_{i=1}^{m} n_i^2/N - \sigma_e^2\right]$$

Noting that $\sum_{i=1}^{m} n_i = N$, this reduces to

$$E(M_1) = \frac{1}{m-1}\left[\sigma_g^2 \left(N - \frac{1}{N}\sum_{i=1}^{m} n_i^2\right) + \sigma_e^2 (m-1)\right]$$

$$= \sigma_e^2 + \frac{1}{m-1}\left[N - \frac{1}{N}\sum_{i=1}^{m} n_i^2\right]\sigma_g^2 \qquad (24)$$

The coefficient of σ_g^2 in (24) is of the same form as given by Snedocor (p.234, 1948).

The within group mean square is computed as

$$M_2 = \frac{1}{N-m}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_1} Y_{ij}^2 - \sum_{i=1}^{m} G_i^2/n_i\right]$$

$$= \frac{1}{N-m}\left[(Y_{11}^2 + Y_{12}^2 + \cdots Y_{1n_1}^2 + Y_{21}^2 + Y_{22}^2 + \cdots Y_{2n_2}^2 + Y_{m1}^2 + Y_{m2}^2 + \cdots Y_{mn_m}^2) - (\frac{G_1^2}{n_1} + \frac{G_2^2}{n_2} + \cdots \frac{G_m^2}{n_m})\right]$$

Taking the expectation term by term we have

$$E(M_2) = \frac{1}{N-m}\left[EY_{11}^2 + EY_{12}^2 + \cdots EY_{1n_1}^2 + EY_{21}^2 + EY_{22}^2 + \cdots EY_{2n_2}^2 + EY_{m1}^2 + EY_{m2}^2 + \cdots EY_{mn_m}^2 - EG_1^2/n_1 - EG_2^2/n_2 - \cdots - EG_m^2/n_m\right] \tag{25}$$

The expectation of the square of any single Y is in no way affected by the number of individuals observed in each group. Therefore, it is given by (14). Substituting in (25) in terms of (14) and (22) we obtain,

$$E(M_2) = \frac{1}{N-m}\left[N\mu^2 + N\sigma_g^2 + N\sigma_e^2 - \sum_{i=1}^{m} n_i\,\mu - \sum_{i=1}^{m} n_i\,\sigma_g^2 - m\sigma_e^2\right]$$

Remembering that μ and σ_g^2 are constants and that $\sum_{i=1}^{m} n_i = N$, this reduces to

$$E(M_2) = \frac{1}{N-m}\left[(N-m)\,\sigma_e^2\right] = \sigma_e^2 \tag{26}$$

Referring to (24) and (26) the analysis of variance with mean square expectations can now be written as follows:

Variance Source	d.f.	Expectations of m.s.
Groups	m-1	$\sigma_e^2 + n' \sigma_g^2$
Within groups	N-m	σ_e^2
Total	N-1	

where $n' = \frac{1}{m-1}\left[N - \frac{\sum_{i=1}^{m} n_i^2}{N}\right]$

General Procedures

Before turning to other examples it will be useful to summarize the general procedures demonstrated in the foregoing examples. Steps in the procedure are listed below.

1. Specification of the model. This includes a symbolic statement of the composition of the individual values that make up the data, assumptions as to whether the various effects are fixed or random, and assumptions concerning whether separate effects vary independently.
2. The composition of each mean square is written out in terms of the model and the steps followed in computing the mean square.
3. The expectation of the mean square is developed term by term.

Rules employed in step 3 may be summarized as follows.

1. The expectation of a constant is the constant itself.
2. The expectation of a variable is the population mean of the variable.
3. The expectation of the square of a variable that has population mean zero is the population variance of the variable.
4. The expectation of the product of a constant and a variable is the product of the constant and the expectation of the variable.
5. The expectation of the product of two variables that have population mean zero is the population covariance of the variables.
6. The population covariance of any two variable effects is zero whenever the particular two effects contributing to any one measurement in the data may be assumed to be independently drawn from their respective populations.

Two points merit special attention.

1. It is desirable to write the model in terms of a general mean so that all effects will have zero as their population mean. This allows taking advantage of 3 and 5 above.
2. If 6 above is kept in mind a great deal of labor can be saved, in writing out the composition of mean squares in expanded form, by omitting product terms that have expectation zero. For example with this in mind (7) might have been written

$$E\,G_i^2 = E\,n^2\mu^2 + E\,n^2g_i^2 + E\,e_{i1}^2 + E\,e_{i2}^2 + \ldots\ldots E\,e_{in}^2$$

for the case where g_i was consided a random variable.

In the case of more complicated analyses than those considered in the foregoing examples, expressions for the composition of the various mean squares may be very long. Rather than follow the procedure outlined above in just the form demonstated by examples 1 and 2, it is more convenient in these cases to recognize that every mean square can be computed as a linear function of one or more "uncorrected" sums of squares and what is commonly called the correction factor. Thus the expectation of a mean square can be obtained by combining the expectations of uncorrected sums of squares and the correction factor in the same way that the sums of square and correction factor were combined to obtain the mean square. The procedure is to find the expectations of the uncorrected sums of squares that must be computed in the analysis and of the correction factor and then combine these appropriately to obtain the expectations of the mean squares.

Example 3

Consider the analysis of data obtained from comparison of n genetic strains of a particular annual crop in a randomized block design at each of s locations in each of t years. Assume r replications in each location each year and that different land or at least a new randomization is used in successive years at each location. The form of the variance analysis is as follows:

Variance Source	d.f.	m.s.
Locations	s-1	
Years	t-1	
L x Y	(s-1)(t-1)	
Reps in years and locations	st(r-1)	
Strains	n-1	M_1
L x Strains	(s-1)(n-1)	M_2
Y x Strains	(t-1)(n-1)	M_3
L x Y x Strains	(s-1)(t-1)(n-1)	M_4
Strains x reps in L and Y	st(r-1)(n-)	M_5
Total	rstn-1	

The model employed will be as follows:

$$Y_{ijkl} = \mu + g_i + a_j + b_k + (ab)_{jk} + (ga)_{ij} + (gb)_{ik} + (gab)_{ijk} + c_{jkl} + (gc)_{ijkl}$$

where μ is the population mean

g_i is the effect of the i-th strain

a_j is the effect of the j-th location

b_k is the effect of the k-th year

$(ab)_{jk}$ is an effect arising from first order interaction between environment conditions of the j-th location and k-th year

$(ga)_{ij}$ is an effect arising from first order interaction of the i-th strain with the j-th location

$(gb)_{ik}$ is an effect arising from first order interaction of the i-th strain with the k-th year

$(gab)_{ijk}$ is an effect arising from second order interaction of the i-th strain with the j-th location and k-th year

c_{jkl} is the effect of the l-th block at the j-th location in the k-th year as a deviation from the mean for that location and year, and

$(gc)_{ijkl}$ is the effect of the plot to which the i-th strain is assigned in the l-th block in the j-th location and k-th year (strictly speaking it also contains a plot-strain interaction effect and the error of measurement, but only in special cases would it be important to indicate this sub-division in the model).

All effects will be considered random variables with mean zero. This would be appropriate if the objective of the work was to compare the strains for use in locations and years of which those involved in the experiment were a random sample, and if the strains represented a random sample from a population from which other strains might have been taken for comparison. It will also be assumed that all effects vary randomly with respect to each other so that all covariances among pairs of effects are zero. This is an appropriate assumption in consideration of the way work like this is usually conducted. Finally, it will be assumed that

$E(ga)^2_{ij}$ is constant over all values of i and j

$E(gb)^2_{ik}$ is constant over all values of i and k

$E(ab)^2_{jk}$ is constant over all values of j and k

$E(gab)^2_{ijk}$ is constant over all values of i, j, and k

$E\ c^2_{jkl}$ is constant over all values of j, k, and l

$E(gc)^2_{ijkl}$ is constant over all values of i, j, k, and l.

The sense of this is that all individual effects within any one of the six kinds belong to a common population and have the variance of that population as the expectation of their squares. This is an assumption very commonly made in connection with analyses of the type in question, though it may not always be justified.

The letter T with appropriate subscripts is used to symbolize different sums of the Y's. For example,

T = grand total

T_i = sum for the i-th variety (over all locations, years, and blocks)

T_j = sum for the j-th location (over all strains, years, and blocks)

T_{ij} = sum for the i-th strain at the j-th location (over all years and blocks) etc.

Carried to its ultimate this means

$$T_{ijkl} = Y_{ijkl}$$

but Y_{ijkl} will be used instead of T_{ijkl}. The uncorrected sums of squares will be symbolized by S with appropriate subscripts. For example,

$S = T^2/nrst$ = the correction factor.

$S_i = \sum_{i=1}^{n} T_i^2/rst$ = uncorrected sum of squares for strains.

$S_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{s} T_{ij}^2/rt$ = uncorrected sum of squares for strain-location totals,

etc.

The process of obtaining the expectations of the mean squares can be amply illustrated by considering only one mean square, say M_2. It is computed as follows:

$$M_2 = \frac{1}{(n-1(s-1)} \left[S_{ij} - (S_i - S) - (S_j - S) - S\right]$$

$$= \frac{1}{(n-1)(s-1)} \left[S_{ij} - S_i - S_j + S\right]$$

Consequently

$$E(M_2) = \frac{1}{(n-1)(s-1)} \left[E\,S_{ij} - E\,S_i - E\,S_j + E\,S\right] \qquad (27)$$

The S's involved have the following composition

$$S_{ij} = \frac{1}{rt} \sum_i \sum_j T_{ij}^2$$

$$S_i = \frac{1}{rst} \sum_i T_i^2$$

$$S_j = \frac{1}{nrt} \sum_j T_j^2$$

$$S = \frac{1}{nrst} T^2$$

It follows that their expectations are,

$$\left.\begin{aligned} E\,S_{ij} &= \frac{1}{rt}\sum_i\sum_j E\,T_{ij}^2 \\ E\,S_i &= \frac{1}{rst}\sum_i E\,T_i^2 \\ E\,S_j &= \frac{1}{nrt}\sum_j E\,T_j^2 \\ E\,S &= \frac{1}{nrst}E\,T^2 \end{aligned}\right\} \qquad (28)$$

As the basis for obtaining the expectations of the T's we must know their composition. The expectation of the square of any of these T's,

$$\begin{aligned} T_{ij} &= \sum_k\sum_l Y_{ijkl} \\ T_i &= \sum_j\sum_k\sum_l Y_{ijkl} \\ T_j &= \sum_i\sum_k\sum_l Y_{ijkl} \\ T &= \sum_i\sum_j\sum_k\sum_l Y_{ijkl} \end{aligned}$$

Expanding these sums in terms of the model for the analysis we have the following:

$$\begin{aligned} T_{ij} = {} & rt\mu + rtg_i + rta_j + r\sum_k b_k + r\sum_k (ab)_{jk} + rt(ga)_{ij} + r\sum_k (gb)_{ik} \\ & + r\sum_k (gab)_{ijk} + \sum_k\sum_l c_{jkl} + \sum_k\sum_l (gc)_{ijkl} \end{aligned}$$

$$\begin{aligned} T_i = {} & rst\mu + rstg_i + rt\sum_j a_j + rs\sum_k b_k + r\sum_j\sum_k (ab)_{jk} + rt\sum_j (ga)_{ij} \\ & + rs\sum_k (gb)_{ik} + r\sum_j\sum_k (gab)_{ijk} + \sum_j\sum_k\sum_l c_{jkl} + \sum_j\sum_k\sum_l (gc)_{ijkl} \end{aligned}$$

$$T_j = nrt\mu + rt \sum_i g_i + nrta_j + nr \sum_k b_k + nr \sum_k (ab)_{jk} + rt \sum_i (ga)_{ij}$$
$$+ r \sum_i \sum_k (gb)_{ik} + r \sum_i \sum_k (gab)_{ijk} + n \sum_k \sum_l c_{jkl} + \sum_i \sum_k \sum_l (gc)_{ijkl}$$

$$T = nrst\mu + rst \sum_i g_i + nrt \sum_j a_j + nrs \sum_k b_k + nr \sum_j \sum_k (ab)_{jk}$$
$$+ rt \sum_i \sum_j (ga)_{ij} + rs \sum_i \sum_k (gb)_{ik} + r \sum_i \sum_j \sum_k (gab)_{ijk} + n \sum_j \sum_k \sum_l c_{jkl}$$
$$+ \sum_i \sum_j \sum_k \sum_l (gc)_{ijkl}$$

The expectation of any of the square of any of these T's is the sum of the expectations of each term in the square. However, since all covariances among different effects are zero (see statement of model) the expectations of all product terms in the square of any T are also zero. Thus only the expectations of the squares of the separate terms in the above expressions contribute to the expectations we are seeking. These can be written directly from inspection of the terms. For example,

$$E(rt\mu)^2 = r^2t^2\mu^2$$

$$E(rtg_i)^2 = r^2t^2\sigma_g^2 \text{ (because } Eg_i^2 = \sigma_g^2)$$

where σ^2 symbolizes the population variance of the effect indicated by subscript

$$E(r \sum_k b_k)^2 = r^2t\sigma_b^2$$ (because (1) the number of b's in the sum indicated is t, (2) $Eb_k^2 = \sigma_b^2$, and (3) the expectation of the product of two b's is zero)

Proceeding in this way the expectations can be written from the equations for the T's as follows:

$$E\,T_{ij}^2 = r^2t^2\mu^2 + r^2t^2\sigma_g^2 + r^2t^2\sigma_a^2 + r^2t\sigma_b^2 + r^2t\sigma_{ab}^2 + r^2t^2\sigma_{ga}^2 + r^2t\sigma_{gb}^2$$
$$+ r^2t\sigma_{gab}^2 + rt\sigma_c^2 + rt\sigma_{gc}^2 \qquad (29a)$$

$$E\,T_i^2 = r^2s^2t^2\mu^2 + r^2s^2t^2\sigma_g^2 + r^2st^2\sigma_a^2 + r^2s^2t\sigma_b^2 + r^2st\sigma_{ab}^2 + r^2st^2\sigma_{ga}^2 + r^2s^2t\sigma_{gb}^2 + r^2st\sigma_{gab}^2 + rst\sigma_c^2 + rst\sigma_{gc}^2 \quad (29b)$$

$$E\,T_j^2 = n^2r^2t^2\mu^2 + nr^2t^2\sigma_g^2 + n^2r^2t^2\sigma_a^2 + n^2r^2t\sigma_b^2 + n^2r^2t\sigma_{ab}^2 + nr^2t^2\sigma_{ga}^2 + nr^2t\sigma_{gb}^2 + nr^2t\sigma_{gab}^2 + n^2rt\sigma_c^2 + nrt\sigma_{gc}^2 \quad (29c)$$

$$E\,T^2 = n^2r^2s^2t^2\mu^2 + nr^2s^2t^2\sigma_g^2 + n^2r^2st^2\sigma_a^2 + n^2r^2s^2t\sigma_b^2 + n^2r^2st\sigma_{ab}^2 + nr^2st^2\sigma_{ga}^2 + nr^2s^2t\sigma_{gb}^2 + nr^2st\sigma_{gab}^2 + n^2rst\sigma_c^2 + nrst\sigma_{gc}^2 \quad (29d)$$

Note that the first of these expressions is constant no matter which genotype-location sum is in question (this is apparent since neither i nor j appears as a subscript in the right hand side of the expression). The same sort of thing is true for the second and third expressions as well. Therefore, equations (28) can be rewritten as follows:

$$\left.\begin{aligned} E\,S_{ij} &= \frac{1}{rt}\left[ns\,E\,T_{ij}^2\right] \\ E\,S_i &= \frac{1}{rst}\left[n\,E\,T_i^2\right] \\ E\,S_j &= \frac{1}{nrt}\left[s\,E\,T_j^2\right] \\ E\,S &= \frac{1}{nrst}\;E\,T^2 \end{aligned}\right\} \quad (30)$$

The only remaining step is to substitute in (27) in terms of equations (29) and (30). Collecting terms involving a common parameter at the same time that the substitutions are made, we obtain,

$$\begin{aligned} E(M_2) = \Big[&\mu^2\,(nrst - nrst - nrst + nrst) + \sigma_g^2\,(nrst - nrst - rst + rst) \\ &+ \sigma_a^2\,(nrst - nrt - nrst + nrt) + \sigma_b^2\,(nrs - nrs - nrs + nrs) \\ &+ \sigma_{ab}^2\,(nrs - nr - nrs + nr) + \sigma_{ga}^2\,(nrst - nrt - rst + rt) \\ &+ \sigma_{gb}^2\,(nrs - nrs - rs + rs) + \sigma_{gab}^2\,(nrs - nr - rs + r) \\ &+ \sigma_c^2\,(ns - n - ns + n) + \sigma_{gc}^2\,(ns - n - s + 1)\Big]\;\frac{1}{(n-1)(s-1)} \\ = \Big[&rt\,(ns - n - s + 1)\,\sigma_{ga}^2 + r\,(ns - n - s + 1)\,\sigma_{gab}^2 \\ &+ (ns - n - s + 1)\,\sigma_{gc}^2\Big]\;\frac{1}{(n-1)(s-1)} \end{aligned}$$

Since (n-1)(s-1) = (ns - n - s + 1) this reduces further to

$$E(M_2) = rt\sigma^2_{ga} + r\sigma^2_{gab} + \sigma^2_{gc}$$

It is worth noting that the mean square for locations is computed as

$$\frac{1}{s-1} (S_j - S)$$

and the one for strains as

$$\frac{1}{n-1} (S_i - S)$$

Thus the expectations of these mean squares could be quickly obtained in terms of information developed in working out $E(M_2)$.

An important practical angle to note is that as one gains experience in working out mean square expectations various short cuts become apparent (for an example see Crump, Biometrics 1946). However, no attempt will be made to describe such short-cuts and when they can be used, as the novice will run less chance of misapplications if he goes through the full procedure in detail until he perceives short-cuts and their rationale by himself. In doubtful cases it is always best to proceed in a straight-forward manner working through the full procedure described above.

Example 4

On occasion estimates of variance components are required from n-fold classification data in which sub-classes are disproportionate and in which in many instances a portion of the sub-classes are not represented at all in the data. In the case of data available to the animal geneticist for estimation of variances arising from genetic variation or genotype-environment interaction this can almost be said to be the rule rather than the exception.

As a specific example suppose that data are available on the annual milk production of cows that were by different sires and that were members of different herds. It will be assumed that members of any particular sire family may have been scattered through two or more herds but not necessarily all herds. Herd effects will vary due to management practices (and perhaps for other reasons), family effects will vary as a result of genotypic variation among sires, and herd-family interaction effects may be presumed to exist. A rational model on which to base analysis of the data would be as follows:

$$Y_{ijk} = \mu + g_i + a_j + (ga)_{ij} + e_{ijk}$$

where Y_{ijk} is the production of the k-th cow that is by the i-th sire and located in the j-th herd,

g_i is the effect of the genotype of the i-th sire (on production by his daughters),

a_j is the effect of the j-th herd,

$(ga)_{ij}$ is an effect due to interaction between average genotype of the i-th family and the environment to which cows are exposed in the j-th herd, and

e_{ijk} is the deviation in production of the k-th cow from the population average for the i-th family in the j-th herd.

It will be assumed that all effects are random with population mean zero, that all individual effects are random with respect to each other so that the expectation for any product of two effects is zero, and finally that

$E(ga)^2_{ij}$ is constant over all values of i and j

and $E\, e^2_{ijk}$ is constant over all values of i and j

If production were measured in various years a realistic model would include other effects but for the purpose of this example we will assume all records were taken in a single year.

There are various computational approaches that may be taken in the use of such data for estimation of variance components, but one of the easiest that is becoming increasingly popular because of its ease is as follows: In terms of our example, four mean squares would be computed: mean squares for (1) families, (2) herds, (3) herd-family subclasses, and (4) cows within herds. The expectations of the first three of these will be linear functions of the variance of all four of the variables in the model. The fourth will have expectation, σ^2_e. Once computed the four mean squares would be equated to their respective expectations to provide four equations in four unknowns (the variances of the four effects) that would then be solved simultaneously to obtain estimates of the four variances.

We will consider the mean square for subclasses (M_{sc}) in detail. It would be computed as,

$$M_{sc} = \left[\sum_i \sum_j T^2_{ij/n_{ij}} - T^2/N \right] \frac{1}{s-1}$$

where n_{ij} is the number of cows of the i-th family in the k-th herd,

T_{ij} is the sum of production by all cows of the i-th family in the k-th herd,

T is the grand total of production by all cows,

N is the total number of cows,

and s is the number of sub-classes represented by one or more cows.

Obviously,

$$E\,M_{sc} = \frac{1}{s-1}\left[\sum_i \sum_j E(T_{ij}^2/n_{ij}) - \frac{1}{N} E\,T^2\right] \tag{31}$$

$$T_{ij} = \sum_{k=1}^{n_{ij}} Y_{ijk} = n_{ij}\,\mu + n_{ij}\,g_i + n_{ij}\,a_j + n_{ij}(ga)_{ij} + \sum_{k=1}^{n_{ij}} e_{ijk}$$

Proceeding in accord with arguments presented in connection with the previous example we can write directly

$$E\,T_{ij}^2 = n_{ij}^2\,\mu^2 + n_{ij}^2\,\sigma_g^2 + n_{ij}^2\,\sigma_a^2 + n_{ij}^2\,\sigma_{ga}^2 + n_{ij}\,\sigma_e^2 \tag{32}$$

In contrast to example 3 this is not constant for all T_{ij} but varies with n_{ij}. We must now find the expectation of T^2.

$$T = \sum_i \sum_j T_{ij} = N\mu + \sum_i n_i g_i + \sum_j n_j a_j + \sum_i \sum_j n_{ij}(ga)_{ij} + \sum_i \sum_j \sum_k e_{ijk}$$

where n_i = total number of cows in the i-th family,

and n_j = total number of cows in the j-th herd.

$$E\,T^2 = N^2\mu^2 + \sum_i n_i^2\,\sigma_g^2 + \sum_j n_j^2\,\sigma_a^2 + \sum_i \sum_j n_{ij}^2\,\sigma_{ga}^2 + N\sigma_e^2 \tag{33}$$

As an example of the detail involved in writing $E\,T^2$ from the expression for T, consider the term, $\sum_i n_i\,g_i$.

$$\sum_i n_i\,g_i = n_1\,g_1 + n_2\,g_2 + \cdots\cdots n_f\,g_f$$

where f is the number of families

$$\left(\sum_i n_i\, g_i\right)^2 = n_1^2\, g_1^2 + n_2^2\, g_2^2 + \ldots. n_f^2\, g_f^2$$

+ product terms that need not be written out since all have zero expectation.

Then $E\left(\sum_i n_i\, g_i\right)^2 = n_1^2\, E\, g_1^2 + n_2^2\, E\, g_2^2 + \ldots. n_f^2\, E\, g_f^2 = \sigma_g^2\,(n_1^2 + n_2^2 + \ldots. n_f^2)$

$= \sigma_g^2 \sum_i n_1^2$ since the expectation of the square of any random g is σ_g^2.

Substituting in (31) in terms of (32) and (33) we obtain,

$$E\, M_{sc} = \frac{1}{s-1}\left[\sum_i \sum_j (n_{ij}\mu^2 + n_{ij}\sigma_g^2 + n_{ij}\sigma_a^2 + n_{ij}\sigma_{ga}^2 + \sigma_e^2)\right.$$

$$\left. - (N\mu^2 + \sigma_g^2 \frac{\sum_i n_i^2}{N} + \sigma_a^2 \frac{\sum_j n_j^2}{N} + \sigma_{ga}^2 \frac{\sum_i \sum_j n_{ij}^2}{N} + \sigma_e^2)\right]$$

$$= \frac{1}{s-1}\left[(N\mu^2 + \sigma_g^2 \sum_i \sum_j n_{ij} + \sigma_a^2 \sum_i \sum_j n_{ij} + \sigma_{ga}^2 \sum_i \sum_j n_{ij} + s\sigma_e^2\right.$$

$$\left. - (N\mu^2 + \sigma_g^2 \frac{\sum_i n_i^2}{N} + \sigma_a^2 \frac{\sum_j n_j^2}{N} + \sigma_{ga}^2 \frac{\sum_i \sum_j n_{ij}^2}{N} + \sigma_e^2)\right]$$

$$= \frac{1}{s-1}\left[\sigma_g^2 \left(\sum_i \sum_j n_{ij} - \frac{\sum_i n_i^2}{N}\right) + \sigma_a^2 \left(\sum_i \sum_j n_{ij} - \frac{\sum_j n_j^2}{N}\right)\right.$$

$$\left. + \sigma_{ga}^2 \left(\sum_i \sum_j n_{ij} - \frac{\sum_i \sum_j n_{ij}^2}{N}\right)\right] + \sigma_e^2$$

Expectations of the other mean squares are obtained by the same procedure as that used for $E\, M_{sc}$. For any particular body of data N, the n_i, the n_j, and the n_{ij}, can be obtained by mere counting and hence, the coefficients of the several variances in $E\, M_{sc}$ can be computed.

Final Comments

The essence of working out mean square expectations can be summarized as follows:

1. It is necessary to know what is meant by expectation.
2. It is necessary to know the values that the definition of expectation imposes on the expectations of (a) a constant (b) a random variate (c) the product of a constant and a random variate (d) the square of a random variate, and (e) the product of two random variates (only the cases of random variates with population mean zero are of special importance).
3. Fundamentally, the procedure is to write the mean square out symbolically in a form that is expanded to the point that it is a linear function of only terms of the type (a) to (e) of point 2 above.
4. When this has been done, knowledge specified in point 2 above, together with the rule that the expectation of a linear function is equal to the same function of the expectations of the separate terms of the quantity for which the expectation is desired provides the basis for writing the desired expectation.
5. From the practical point of view, many of the steps can and will be performed only mentally (will not be written out). However, in case of doubt, writing steps out in detail is likely to insure against an occasional serious error. There are rules-of-thumb that can sometimes be used but their application involves risk of error unless the entire matter is so well understood that the reason why these rules work in specific cases is entirely clear. Otherwise they may be applied in cases where they do not work.

For supplementary reading on the derivation of mean square expectations see Anderson and Bancroft (1952) and Kempthorne (1952).

Literature Cited

Anderson, R. L. and T. A. Bancroft (1952) Statistical Theory in Research, McGraw-Hill. New York.

Crump, S. Lee (1946) The Estimation of Variance Components in Analysis of Variance. Biometrics Bull. 2:7-11.

Kempthorne, Oscar (1952) The Design and Analysis of Experiments. John Wiley and Sons, Inc. New York.