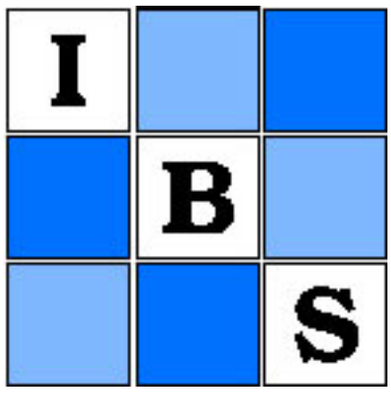

THE COMBINATION OF ESTIMATES FROM DIFFERENT EXPERIMENTS*

WILLIAM G. COCHRAN
The Johns Hopkins University,
Baltimore, Maryland

1. INTRODUCTION

When we are trying to make the best estimate of some quantity μ that is available from the research conducted to date, the problem of combining results from different experiments is encountered. The problem is often troublesome, particularly if the individual estimates were made by different workers using different procedures. This paper discusses one of the simpler aspects of the problem, in which there is sufficient uniformity of experimental methods so that the ith experiment provides an estimate x_i of μ, and an estimate s_i of the standard error of x_i . The experiments may be, for example, determinations of a physical or astronomical constant by different scientists, or bioassays carried out in different laboratories, or agricultural field experiments laid out in different parts of a region. The quantity x_i may be a simple mean of the observations, as in a physical determination, or the difference between the means of two treatments, as in a comparative experiment, or a median lethal dose, or a regression coefficient.

The problem of making a combined estimate has been discussed previously by Cochran (1937) and Yates and Cochran (1938) for agricultural experiments, and by Bliss (1952) for bioassays in different laboratories. The last two papers give recommendations for the practical worker. My purposes in treating the subject again are to discuss it in more general terms, to take account of some recent theoretical research, and, I hope, to bring the practical recommendations to the attention of some biologists who are not acquainted with the previous papers.

The basic issue with which this paper deals is as follows. The simplest method of combining estimates made in a number of different experiments is to take the arithmetic mean of the estimates. If, however, the experiments vary in size, or appear to be of different precision, the investigator may wonder whether some kind of weighted mean would be more precise. This paper gives recommendations about the kinds of weighted mean that are appropriate, the situations in which they

*Department of Biostatistics, Paper No. 292. This work was assisted by a contract with the Office of Naval Research.

are appropriate, and the circumstances in which the unweighted mean is to be preferred. Methods for obtaining a standard error to be attached to the final estimate are also presented.

The mathematical theory which bears on the problem is complex, and some of the recommendations are based on approximations in theory. Wherever possible, the recommendations are documented by references to published papers. Some theoretical issues are discussed briefly in sections 6 to 9 in cases where the documentation available in the literature does not seem adequate.

2. MATHEMATICAL MODELS

As will appear later, the best combined estimate depends on the nature of the data. It is advisable to consider the preliminary question:

Do the values of the x_i agree among themselves within the limits of their experimental errors?

If they do, we may postulate an underlying mathematical model of the form

$$x_i = \mu + e_i \,, \tag{1}$$

where e_i is the experimental error of x_i .

If the values of the x_i differ by more than can be accounted for by their experimental errors, we require a model of the form

$$x_i = \mu_i + e_i \,, \tag{2}$$

where μ_i , which might be called the "true value" in the ith experiment, varies from one experiment to another. There are numerous reasons why such variations may exist. They may be the result of differences in the experimental techniques used in the different experiments, of biases that vary in size from one experiment to another, or of real changes in μ_i due to the environment in which the experiment is conducted. Frequently the investigator is able to predict, from general knowledge or from past experience with the same type of data, whether the μ_i are likely to vary. In agricultural experiments that are located on farmers' fields throughout an area, for instance, it is quite commonly found that the response to a fertilizer exhibits a real variation from field to field. This variation is often described as an *interaction* of the effect with experiments. Tests of significance for this interaction will be presented later.

If an interaction exists, the type of combined estimate that is wanted requires careful consideration. It is necessary to take into account the purpose for which the combined estimate is to be made and the reasons for the presence of interaction, in so far as these can be discovered.

The following are illustrations of some of the situations that may arise.

(1) In the determination of a physical constant, we might conclude that interaction exists because some of the experiments (e.g. the earlier ones) were done by a technique that is subject to a bias of unknown magnitude, whereas the remainder of the experiments appear to be unbiased. In this event we would presumably discard the results from the biased experiments and consider only a combination of results from the unbiased experiments.

(2) In agricultural experiments the variation in μ_i may be due mainly to the soil type on which the experiments are conducted. The experiments can then be classified into groups, each of which represents a specific soil type. It may also happen that the number of experiments on a given soil type is not at all proportional to the area of the crop under that soil type in practical farming, perhaps because the experiments were deliberately set up to include some of the rarer types. In this case, if our object is to estimate a mean over some defined farming area, we might adopt the kind of weighted mean that is appropriate to stratified sampling. Thus if $\bar{x}_j$ is the estimated mean for the jth soil type, and A_j is the estimated area of the crop under this type in the population, the overall mean is taken as $\sum A_j\bar{x}_j/\sum A_j$.

(3) In the preceding situation we might decide, alternatively, not to estimate the overall mean at all, but to present the individual estimates for the different groups or strata. This practice is advisable where the μ_i vary so much that different practical recommendations must be given in different strata. Of course, such recommendations are feasible only when the user of the results knows to which stratum he belongs. An example might be experiments on the feeding of chickens, where the results vary with the breed of the chickens.

(4) Occasionally, in laboratory experiments which were thought to be well-controlled, large interactions may appear for which no adequate explanation can be given. In this event it might be best to hand the problem back to the experimenters, on the grounds that there is not much point in attempting a "best" combined estimate until the experimenters can reach better agreement in their results, or at least find out why they disagree.

These illustrations, which do not exhaust the possibilities, bring out the point that the combination of the individual estimates is not a routine matter, but requires clear thinking about both the nature of the data and the function of a combined estimate. However, unless it is decided that no type of combined estimate will serve a useful purpose, we do face the problem of combining at least over certain subgroups of the experiments.

In the remainder of this paper it will be assumed that the experiments which we have decided to combine are a random sample from the population of experiments about which we wish information. This assumption is far from being universally true in practice and should be examined before adopting the methods in this paper, since series of experiments often come into existence in a rather haphazard way.

The discussion will deal only with the combination of a single estimate x_i from each experiment. When each experiment contains more than 2 treatments, we may wish to make a combined analysis of all the experimental results. Some methods for handling this problem are given by Yates and Cochran (1938), Cochran and Cox (1950) and Kempthorne (1952).

3. EXPERIMENTS OF THE SAME SIZE AND THE SAME PRECISION

The simplest case is that in which all k experiments are of exactly the same type, with no missing data, and the estimates x_i all have the same error variance σ^2. In this event the estimated variances s_i^2 will each have n degrees of freedom and will each be unbiased estimates of σ^2. To avoid confusion, note that the symbols s_i^2 and σ^2 refer to the variance of x_i , not to the variance per single observation in the experiment.

This case will occur when every experiment has the same precision per observation, and x_i is the same linear function of the observations in the experiment. Thus the variance σ^2 will be of the form σ_0^2/f, where σ_0^2 is the common variance per observation, and f is a divisor which is the same in all experiments. For example, if x_i is an unweighted mean over r replications, $f = r$, and if x_i is the difference between two such means, $f = r/2$. This case would not apply, however, if x_i were the regression of yield on plant number, because the variance of x_i would depend on the distribution of plant numbers in the ith experiment.

This case can be handled by familiar and elementary methods, but is included for completeness.

To test whether the x_i are of the same precision we may apply Bartlett's test, in which we compute χ^2, with $(k - 1)$ degrees of freedom, as

$$\chi^2 = \frac{2.303}{C}\left[nk \log \bar{s}^2 - \sum_{i=1}^{k} n \log s_i^2\right] \tag{3}$$

where $\bar{s}^2$ is the arithmetic mean of the s_i^2 and

$$C = 1 + \frac{(k + 1)}{3nk}$$

Although the investigator can never be sure that the x_i all have the same variance, it is suggested, as a working rule, that the methods in

this section are adequate whenever Bartlett's χ^2 is not significant at the 5 per cent level. This opinion is based on the results of a number of sets of data which were worked with and without the assumption of homogeneity. Methods which do not require the assumption are given in section 5.

On the assumption that the s_i^2 are homogeneous, the interaction of the x_i with experiments can be tested by means of a standard F-test in the analysis of variance (table 1).

TABLE 1
TEST OF THE VARIATION IN x_i FROM EXPERIMENT TO EXPERIMENT

Source of variation	d.f.	Mean Squares
Interaction with experiments	$(k-1)$	$s_b^2 = \Sigma(x_i - \bar{x})^2/(k-1)$
Pooled internal error	nk	$\bar{s}^2$

Interactions Absent.

If there is no interaction (μ_i all equal to μ), then from equation (1) each x_i is an estimate of μ with common variance σ^2. Hence, if the x_i are approximately normally distributed, the recommended estimate of μ is their unweighted mean $\bar{x}$, with variance σ^2/k.

To find a sample estimate of the standard error of $\bar{x}$, we may note that the quantities $\bar{s}^2$ and s_b^2 are both estimates of σ^2, with nk and $(k-1)$ degrees of freedom, respectively. The best estimate of the standard error of $\bar{x}$ is the pooled value

$$\text{s.e.}(\bar{x}) = \sqrt{\frac{nk\bar{s}^2 + (k-1)s_b^2}{k(nk+k-1)}} \tag{4}$$

with $(nk + k - 1)$ degrees of freedom. Since nk is usually much greater than $(k-1)$, the use of $\bar{s}/\sqrt{k}$ as the estimated standard error is not uncommon.

Interactions Present.

In this event, the quantity to be estimated is the population mean μ of the μ_i . Let σ_μ be the standard deviation of the distribution of the μ_i . Then from equation (2),

$$x_i = \mu_i + e_i = \mu + (\mu_i - \mu) + e_i$$

It follows that the estimates x_i vary about μ with variance $(\sigma_\mu^2 + \sigma^2)$.

Since the x_i are still of equal precision as estimates of μ, the un-

weighted mean $\bar{x}$ is still the best estimate of μ. However, the variance of $\bar{x}$ is now $(\sigma_\mu^2 + \sigma^2)/k$, and expression (4) cannot be used for the estimated standard error of $\bar{x}$. It is easy to show algebraically that s_b^2 in table 1 is an unbiased estimate of $(\sigma_\mu^2 + \sigma^2)$, so that for the standard error of $\bar{x}$ we use

$$\text{s.e.}(\bar{x}) = \frac{s_b}{\sqrt{k}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{k(k-1)}} \tag{5}$$

To summarize, the only decision that needs to be made is whether we will regard interactions as present or absent. The more conservative procedure is always to regard interactions as present, since estimate (5) for the standard error is valid whether interactions are present or not. If the number of experiments, k, is small, however, this estimate has low precision, and one is tempted to use the pooled estimate (4).

A procedure followed by some workers is to pool whenever the F from table 1 is not significant at the 5 per cent level. This has been criticized by others on the grounds that it may underestimate the standard error of $\bar{x}$. The consequences of a rule of this kind have been examined by Bancroft (1944) and Paull (1950) for a series of values of the ratio $I = \sigma_\mu^2/\sigma^2$. Their results show that the rule is somewhat hazardous, in that for moderate values of I (say between 1/4 and 4) it underestimates the standard error and gives too many significant results in a subsequent t-test of $\bar{x}$. The alternative rule of pooling only when $F < 2$, suggested by Paull, is safer: its chief defect is that if I is small, it may slightly overestimate the standard error.

4. EXPERIMENTS OF DIFFERENT SIZES BUT OF THE SAME PRECISION PER OBSERVATION

Sometimes the experiments that are being combined differ in size and structure, but there is reason to believe that experimental error variances *per observation* are the same in all experiments. If σ_0^2 denotes this common variance, the variance σ_i^2 of the estimate x_i will be of the form $\sigma_i^2 = \sigma_0^2/f_i$, where f_i is a factor depending on the type of experiment. For instance, if x_i is a mean over r_i replications, $f_i = r_i$, and if x_i is the difference between two such means, $f_i = r_i/2$. Similarly, if s_{0i}^2 denotes the estimated error variance per observation in the ith experiment, $s_i^2 = s_{0i}^2/f_i$.

Example 1.

This example was obtained by selecting two treatments from a series of experiments on the effectiveness of carbon tetrachloride in

killing worms (Nippostrongylus muris) which are parasitic on rats (Whitlock and Bliss, 1943). Each rat was injected with 500 larvae. Eight days later, the rats were treated with varying doses of CCl_4 and two days later all rats were killed and the numbers of adult worms were counted for each rat. The treatments to be discussed are the control (no CCl_4) and a dose of 0.063 cc per rat. Three experiments included both treatments, with numbers of replications as follows.

Expt.	Control	0.063 cc CCl_4
1	5	3
2	5	5
3	6	7

The relevant data are shown in table 2.

TABLE 2
ESTIMATES (x_i) AND VARIANCES PER RAT (s^2_{0i})

Expt.	No. of adult worms		Difference			d.f.
	Control	CCl_4	x_i	s^2_{0i}	f_i	n_i
1	290.4	204.0	86.4	3,223	1.875	10
2	323.2	165.2	158.0	8,370	2.500	14
3	274.0	262.7	11.3	2,606	3.231	16

The estimates to be combined are the differences x_i between the mean recoveries for the control and the treated rats. The values of f_i may be verified from the numbers of replications already reported. In computing the variances per rat, s^2_{0i}, the data from treatments with smaller doses of CCl_4 were also used, so that the degrees of freedom are larger than would be provided by the two treatments discussed here.

The first step is to apply Bartlett's χ^2 test to the estimated variances per observation.

$$\chi^2 = \frac{2.303}{C} [n_c \log \bar{s}^2_0 - \sum n_i \log s^2_{0i}] \qquad (6)$$

where

$$n_c = \sum n_i = 40$$

$$\bar{s}_0^2 = \sum n_i s_{0i}^2/n_c = (191{,}106)/40 = 4{,}777.6$$

$$C = 1 + \frac{1}{3(k-1)}\left\{\sum \frac{1}{n_i} - \frac{1}{n_c}\right\} = 1.035.$$

The value of χ^2 is 5.59, with $k - 1 = 2$ degrees of freedom. The significance probability is about 0.06, and it is doubtful whether the variances per rat can be considered homogeneous. For the present, this assumption will be made: in section 5 the example will be re-worked without this assumption.

In order to test whether the estimates x_i agree with each other within the limits of their experimental error variances, we carry out a conventional analysis of variance on a single-observation basis (table 3).

TABLE 3
ANALYSIS OF VARIANCE ON A SINGLE-OBSERVATION BASIS

Source of variation	d.f.	Sum of squares	Mean squares
Interaction with expts.	$(k-1)$	$\Sigma f_i(x_i - \bar{x}_w)^2$	$s_{0b}^2 = \Sigma f_i(x_i - \bar{x}_w)^2/(k-1)$
Pooled error	n_c	$\Sigma n_i s_{0i}^2$	$\bar{s}_0^2 = \Sigma n_i s_{0i}^2/n_c$

Note that the factors f_i are used as weights in computing the sum of squares for the interaction with experiments.
The quantity $\bar{x}_w$ in table 3 is the *weighted* mean

$$\bar{x}_w = \sum f_i x_i / \sum f_i \qquad (7)$$

The F-ratio, $s_{0b}^2/\bar{s}_0^2$, gives a test of significance of the presence of interactions.

At this point there are three situations to be considered.

Interactions absent.

This case is a familiar one in elementary text-books. We revert to the mathematical model

$$x_i = \mu + e_i$$

so that x_i is an estimate of μ with variance σ_0^2/f_i. By least squares theory, the best combined estimate of μ is the weighted mean $\bar{x}_w$, and its variance is

$$V(\bar{x}_w) = \sigma_0^2/\sum f_i$$

The most precise combined estimate of σ_0^2 is obtained by pooling the sums of squares in table 3 to give

$$\hat{s}_0^2 = \frac{\sum f_i(x_i - \bar{x}_w)^2 + \sum n_i s_{0i}^2}{k - 1 + n_c}$$

The standard error of $\bar{x}_w$ is taken as $\hat{s}_0/\sqrt{\sum f_i}$, with $(k - 1 + n_c)$ degrees of freedom.

Interactions large.

With the more general model

$$x_i = \mu + (\mu_i - \mu) + e_i ,$$

the variance of x_i, as an estimate of μ, is

$$V(x_i) = \sigma_\mu^2 + \frac{\sigma_0^2}{f_i}$$

If the values of σ_μ^2 and σ_0^2 were known, the least squares estimate of μ would be the *semi-weighted* mean

$$\bar{x}_{sw} = \sum W_i x_i / \sum W_i \qquad \text{where} \qquad W_i = \frac{1}{\sigma_\mu^2 + \frac{\sigma_0^2}{f_i}} \tag{8}$$

The semi-weighted mean (8) includes the weighted mean (7) as a particular case, since it reduces to the weighted mean when $\sigma_\mu = 0$. At the other extreme, when interactions are large, σ_μ is large relative to σ_0 and the semi-weights are all approximately equal. The semi-weighted mean then differs little from the unweighted mean.

Since it is not profitable to go to the extra trouble of computing the semi-weighted mean unless we are confident that there will be a worthwhile gain in precision over the unweighted mean, the precision of the unweighted mean is compared with that of the semi-weighted mean in section 6. The relative precision is found to depend on two factors:

(i) The ratio I of the interaction variance σ_μ^2 to the average of the experimental error variances σ_0^2/f_i. The higher the value of I, the smaller is the loss of precision resulting from the use of the unweighted mean.

(ii) The amount of variation in the factors f_i. As the variation in the f_i increases, the loss of precision resulting from the unweighted mean increases.

The theoretical examination in section 6 leads to the rules given in table 4.

TABLE 4

WORKING RULES FOR THE USE OF THE UNWEIGHTED MEAN

If ratio of largest to smallest f_i	Use the unweighted mean whenever
<2	$F > 3$
between 2 and 6	$F > 4$
>6	$F > 5$

The rules will be illustrated from example 1. The analysis of variance appears in table 5.

TABLE 5

ANALYSIS OF VARIANCE FOR DATA IN TABLE 2

Source of variation	d.f.	Sums of squares	Mean squares	F
Interaction with expts.	2	30,506	$s_{0b}^2 = 15,253$	3.19
Pooled error	40	191,106	$s_0^2 = 4,778$	

The F-ratio, 3.19, is almost at the 5 per cent level, indicating a variation in the effectiveness of the CCl_4 from experiment to experiment. From table 2 we see that the ratio of the largest to the smallest f_i is less than 2. By the rule in table 4, the unweighted mean of the x_i is recommended since F is over 3. The estimate is

$$\bar{x} = \frac{86.4 + 158.0 + 11.3}{3} = 85.2$$

The standard error of $\bar{x}$ is given by

$$\text{s.e.}_{\bar{x}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{k(k-1)}} = \sqrt{\frac{10,763}{(3)(2)}} = 42.3$$

The usual procedure is to attribute $(k - 1)$ or 2 degrees of freedom to this standard error. This is not quite correct, because the x_i are presumably not all of the same precision as estimates of μ, even though we have decided to use their unweighted mean as an overall estimate.

An approximate adjustment which gives a smaller number of degrees of freedom is developed in section 7, although my experience is that it is rarely needed for this application. The adjustment requires some

supplementary calculations which are given for another purpose in the subsequent table 6.

In column 2 of table 6, let

$$v_i = s_\mu^2 + \frac{\bar{s}_0^2}{f_i}$$

and let

$$V_1 = \frac{\sum v_i}{k} = 6{,}211 \; : \; V_2 = \frac{\sum v_i^2}{k} = 38{,}773{,}457.$$

The adjusted number of degrees of freedom is

$$n_e = \left\{ \frac{(k-1)^2 V_1^2}{(k-2)V_2 + V_1^2} \right\} = 1.99 \tag{9}$$

The adjustment has a negligible effect.

Interactions moderate.

With most sets of data, either the weighted or the unweighted mean will prove to be satisfactory. There remain some cases in which, although F is not large, we believe from the nature of the data that interactions are likely to be present, and are reluctant to rely on either the unweighted or the weighted mean. A sample semi-weighted mean, which is an analogue of the semi-weighted mean in equation (8), may be tried.

In equation (8) the true semi-weights W_i were given as

$$W_i = \frac{1}{\sigma_\mu^2 + \frac{\sigma_0^2}{f_i}}$$

The first step is to obtain sample estimates of σ_μ^2 and σ_0^2 .

It is easily shown by algebra that the expectations of the two mean squares in the analysis of variance in table 3 are as follows.

$$E(s_{0b}^2) = \sigma_0^2 + \bar{f}'\sigma_\mu^2$$
$$E(\bar{s}_0^2) = \sigma_0^2$$

where

$$\bar{f}' = \frac{1}{(k-1)} \left\{ \sum f_i - \frac{\sum f_i^2}{\sum f_i} \right\}$$

The quantity $\bar{f}'$ is always smaller than the arithmetic mean of the f_i . From these results, an unbiased estimate of σ_μ^2 is

$$s_\mu^2 = (s_{0b}^2 - \bar{s}_0^2)/\bar{f}' \tag{10}$$

Finally, for the sample semi-weighted mean we take

$$\bar{x}_{sw} = \sum \hat{W}_i x_i / \sum \hat{W}_i$$

where

$$\hat{W}_i = \frac{1}{s_\mu^2 + \frac{\bar{s}_0^2}{f_i}}$$

To illustrate the method from example 1, the analysis of variance (table 5) gives

$$s_{0b}^2 = 15{,}253 : \bar{s}_0^2 = 4{,}778.$$

The value of $\bar{f}'$ is found from table 2 to be

$$\bar{f}' = \frac{1}{2}\left\{7.606 - \frac{20.205}{7.606}\right\} = 2.475$$

TABLE 6
CALCULATION OF SEMI-WEIGHTED MEAN

Expt.	$s_\mu^2 + \frac{\bar{s}_0^2}{f_i} = 4{,}232 + \frac{4{,}778}{f_i}$	Reciprocal $\hat{W}_i$	x_i
1	6,780	.000147	86.4
2	6,143	.000163	158.0
3	5,711	.000175	11.3
Total		.000485	

Hence, from equation (10),

$$s_\mu^2 = \frac{15{,}253 - 4{,}778}{2.475} = 4{,}232.$$

The semi-weights are computed in table 6.

$$\bar{x}_{sw} = \frac{(86.4)(147) + (158.0)(163) + (11.3)(175)}{485} = 83.4$$

This does not differ materially from the unweighted mean, 85.2.

The standard error of $\bar{x}_{sw}$ is, approximately,

$$\text{s.e.}(\bar{x}_{sw}) = \frac{1}{\sqrt{\sum \hat{W}_i}} = \frac{1}{\sqrt{.000485}} = 45.4$$

This formula may give values that are slightly too low, since it ignores the fact that the weights $\hat{W}_i$ are subject to sampling errors. For t-tests, it is suggested that $(k - 1)$ degrees of freedom be assigned to the standard error, although the distributions involved have not yet been adequately investigated.

5. EXPERIMENTS OF UNEQUAL PRECISION PER OBSERVATION

The methods in this section are to be used when Bartlett's χ^2 is significant or when, for any other reason, the investigator does not wish to assume that the variances per observation are equal. The methods are the same whether the experiments are identical in size and type or not.

As before, the estimate in the ith experiment is denoted by x_i, and s_i^2 is an unbiased estimate of the error variance of x_i, based on n_i degrees of freedom.

The test for interactions.

The first step is to test for the presence of interactions. One approach is to calculate the ordinary mean square deviation of the x_i, i.e.

$$\frac{\sum (x_i - \bar{x})^2}{k - 1}$$

If there are no interactions, this quantity is an unbiased estimate of

$$\bar{\sigma}^2 = \frac{\sum \sigma_i^2}{k}$$

Consequently, an approximate F-test of the interactions is made from the ratio

$$F = \frac{\sum (x_i - \bar{x})^2}{(k - 1)\bar{s}^2}$$

where $\bar{s}^2 = \sum s_i^2/k$.

From an elementary point of view, the degrees of freedom might be taken as $(k - 1)$ and $n_e = \sum n_i$. Actually, the calculated F does not follow the tabular F-distribution, because the x_i vary in precision. The tabular F-distribution might still be used, as an approximation, by reducing the numbers of degrees of freedom ascribed to F. Following equation (9), the degrees of freedom in the numerator may be taken as

$$\nu_1 = \frac{(k - 1)^2 V_1^2}{(k - 2)V_2 + V_1^2} \tag{11}$$

where

$$V_1 = \frac{\sum s_i^2}{k} : V_2 = \frac{\sum s_i^4}{k}$$

For the denominator, a familiar rule is

$$\nu_2 = \frac{(\sum s_i^2)^2}{\sum \frac{s_i^4}{n_i}} \tag{12}$$

There is, however, the further objection that F may be insensitive in detecting the presence of interactions, because experiments with low precision receive the same weight as those with high precision.

An alternative course, less open to these objections, is to base the test of interactions on the *weighted* sum of squares of deviations

$$Q = \sum_{i=1}^{k} w_i(x_i - \bar{x}_w)^2,$$

where

$$w_i = 1/s_i^2$$

$$\bar{x}_w = \sum w_i x_i / \sum w_i$$

If the degrees of freedom n_i are large, Q follows the χ^2 distribution with $(k - 1)$ degrees of freedom. For moderate values of n_i , adjustments which transform Q so as to give a better approximation to χ^2 have been worked out by Cochran (1937) for n_i all equal, and by James (1951). Welch (1951) transforms Q so that it may be referred to the F-table: his test and that of James are similar in that both neglect terms of order $1/n_i^2$. Although the range of applicability of these tests, which are all approximations, is not yet known, it is suggested that they be used down to $n_i = 6$. To perform Welch's test, we compute the auxiliary quantity

$$a = \sum_{i=1}^{k} \frac{1}{n_i}\left(1 - \frac{w_i}{w}\right)^2$$

where $w = \sum w_i$. Then

$$F_w = \frac{Q}{(k-1) + \frac{2(k-2)a}{(k+1)}}$$

with degrees of freedom

$$\nu_1 = (k - 1) : \nu_2 = \frac{(k^2 - 1)}{3a}$$

The F and F_w tests will be illustrated by the data in table 2 for example 1. Although this example was analysed under the assumption that the experiments were of equal precision per observation, the probability value for Bartlett's χ^2, 0.06, casts doubt on this assumption. The first step is to compute the quantities $s_i^2 = s_{0i}^2/f_i$: these are the estimated variances of the x_i . The remainder of the calculations are arranged in table 7.

TABLE 7
CALCULATIONS FOR THE TEST OF INTERACTIONS

x_i	s_i^2	$10^2 w_i$	$w_i x_i$	w_i/w	$(1 - w_i/w)^2$	n_i	$(1 - w_i/w)^2/n_i$
86.4	1719	.0582	.05028	.275	.526	10	.0526
158.0	3348	.0299	.04724	.141	.738	14	.0527
11.3	807	.1239	.01400	.584	.173	16	.0108
	5874	.2120	.11152	1.000			$a = .1161$

The F-test.

$$\sum (x_i - \bar{x})^2 = 10{,}763 : \bar{s}^2 = 1{,}958 : F = \frac{5{,}381}{1{,}958} = 2.75$$

The degrees of freedom from an elementary point of view would be 2 and 40. Formulas (11) and (12) will be found to give 1.7 and 30.3 degrees of freedom, respectively. By interpolation between $F(1, 30)$ and $F(2, 30)$, the significance probability comes out at 0.09.

The F_w-test.

$$Q = \sum w_i x_i^2 - (\sum w_i x_i)^2/w = 11.966 - (.11152)^2/(.002120) = 6.10$$

$$F_w = \frac{6.10}{2 + \frac{2(1)(.1161)}{4}} = \frac{6.10}{2.058} = 2.96$$

$$\nu_1 = 2 : \nu_2 = \frac{8}{3(.1161)} = 23.0$$

The significance probability is about 0.07.

The test of interactions is of importance because, as pointed out in section 2, the presence of interactions affects our interpretation of the data and may determine the kind of mean that will be useful. In a borderline case, as in this example, the investigator should take into account both the significance probability and any other knowledge of

the data in deciding whether to regard interactions as present or absent. The conservative decision, when in doubt, is to assume interactions present, since the techniques for this situation remain valid even if interactions are absent.

Experience in the application of the F- and F_w-tests indicates that although the F_w-test is more sensitive, the F-test, which is simpler to compute, is usually adequate for diagnostic purposes. Consequently, the working rules to be given later are based on the value of F.

Interactions absent.

If we are willing to assume that interactions are absent, one method of combination is to weight each x_i inversely as its estimated variance s_i^2, forming the weighted mean

$$\bar{x}_w = \frac{\sum w_i x_i}{w}, \qquad w_i = 1/s_i^2 : w = \sum w_i$$

The standard error of $\bar{x}_w$ is given approximately by a formula due to Meier (1953), with an adjustment by Cochran and Carroll (1953),

$$\text{s.e.}(\bar{x}_w) = \sqrt{\frac{1}{w}\left\{1 + \frac{4}{w^2}\sum \frac{1}{n_i'} w_i(w - w_i)\right\}} \tag{13}$$

where

$$n_i' = n_i - \frac{4(k-2)}{(k-1)}$$

If the n_i are all equal, formula (13) reduces to the slightly simpler expression

$$\text{s.e.}(\bar{x}_w) = \sqrt{\frac{1}{w}\left\{1 + \frac{4}{n'}\left(1 - \frac{\sum w_i^2}{w^2}\right)\right\}} \tag{14}$$

The term inside the brackets is an adjustment which takes account of sampling errors in the weights $1/s_i^2$ as estimates of the true weights $1/\sigma_i^2$, and also of the fact that the principal term inside the square root, $1/w$, tends to be an underestimate of the corresponding population expression. These formulas require $n_i \geq 8$: for values of n_i below 8, see section 8.

For the approximate number of degrees of freedom n_e to be attached to this standard error, Meier (1953) suggests

$$n_e = \frac{w^2}{\sum \frac{w_i^2}{n_i}} \tag{15}$$

If the n_i are small, the sampling errors in the weights may be large enough so that the weighted mean is no more precise than the unweighted mean $\bar{x}$, whose standard error is

$$\text{s.e.}(\bar{x}) = \frac{\sqrt{\sum s_i^2}}{k} \tag{16}$$

The approximate number of degrees of freedom n_e' for this s.e. is

$$n_e' = \frac{(\sum s_i^2)^2}{\sum \left(\frac{s_i^4}{n_i}\right)} \tag{17}$$

In seeking some rule which will help in deciding whether to use $\bar{x}$ or $\bar{x}_w$, it is natural to try to base the rule on the value of Bartlett's χ^2, since this will already have been calculated in many cases. Unfortunately, the relation between this χ^2 and the relative precision of $\bar{x}$ to $\bar{x}_w$ is not simple. When the degrees of freedom n_i are large, χ^2 can detect relatively small differences in precision which make $\bar{x}_w$ only slightly more precise than $\bar{x}$. When the n_i are small, on the other hand, χ^2 may sometimes be non-significant even when $\bar{x}_w$ would be substantially better than $\bar{x}$. As a rough guide to the relative precision R of $\bar{x}$ to $\bar{x}_w$, the following formula is suggested.

$$R \doteq \frac{\bar{n}}{(\bar{n} - 2)} e^{-2\chi^2/n_c} \tag{18}$$

where

$$n_c = \sum n_i : \bar{n} = n_c/k$$

This formula was derived as a mathematical approximation and has been checked on a number of sets of data.

Since $\bar{x}$ is preferable on account of its simplicity unless $\bar{x}_w$ brings a worthwhile gain in precision, the investigator will not go far wrong in using $\bar{x}$ unless R is less than 0.9. My experience with actual data has been that often there is little to choose between $\bar{x}$ and $\bar{x}_w$, but occasionally $\bar{x}_w$ wins handsomely.

A warning given by Yates and Cochran (1938) should be repeated. It sometimes happens that there is a correlation between x_i and s_i^2 , for instance when experiments which have large responses also exhibit high variability. In this event a weighted mean gives too much weight to experiments where the response is low and will be biased.

Example 2 illustrates the rule for choosing between $\bar{x}$ and $\bar{x}_w$.

Example 2.

The data in table 8 are the responses in sugar per acre to an applica-

TABLE 8
RESULTS OF 4 EXPERIMENTS ON SUGAR-BEET

Response to P (cwt) x_i	s_i^2	$w_i = 1/s_i^2$
+1.3	4.973	0.20
+0.4	1.416	0.71
+0.7	6.864	0.15
+2.5	2.958	0.34
Total	16.211	$1.40 = w$

tion of superphosphate in 4 experiments on heavy loam soils in the 1936 series of fertilizer trials on sugar-beet in England. Each experiment provided 15 degrees of freedom for error, giving $n_e = 60$.

The value of Bartlett's χ^2 is 9.27, with a probability of about 0.03. In the test for interactions the value of F is less than 1. The response to P had also shown no sign of interactions in several other sets of these sugar-beet experiments, so that the assumption of negligible interactions appeared justifiable.

By formula (18), the crude estimate of R is

$$R \doteq \frac{\bar{n}}{(\bar{n} - 2)} e^{-2\chi^2/n_e} = \left(\frac{15}{13}\right) e^{-18.54/60} = 0.85$$

The weighted mean is suggested. Its value is

$$\bar{x}_w = \frac{(1.3)(0.20) + (0.4)(0.71) + (0.7)(0.15) + (2.5)(0.34)}{1.40} = 1.07$$

For the standard error, the simpler form in equation (14) can be used.

$$n' = n - \frac{4(k - 2)}{(k - 1)} = 15 - \frac{(4)(2)}{(3)} = 12.3$$

Hence by equation (14),

$$\text{s.e.}(\bar{x}_w) = \sqrt{\frac{1}{1.40}\left\{1 + \frac{4}{12.3}\left(1 - \frac{.6822}{1.96}\right)\right\}} = 0.93$$

From equation (15), the approximate number of degrees of freedom is

$$n_e = \frac{w^2}{\sum \frac{w_i^2}{n_i}} = \frac{n\,w^2}{\sum w_i^2} = \frac{(15)(1.96)}{.6822} = 43$$

The unweighted mean may be verified to be 1.22 ± 1.01.

When the numbers of degrees of freedom in the individual experiments are less than 8, the weighted mean will seldom be more precise than the unweighted mean. With the weighted mean, one or two experiments tend to receive very large weights and almost determine the value of the overall mean. If Bartlett's χ^2 is large, the investigator may still feel that some kind of weighting is desirable. A suggested procedure is *partial weighting* (Yates and Cochran, 1938). The same weight is given to all experiments with relatively low values of s_i^2, this weight being $\bar{w}_p = 1/\bar{s}_p^2$, where $\bar{s}_p^2$ is the mean of the s_i^2 over those experiments that are chosen to have equal weight. Each of the remaining experiments receives its individual weight $w_i = 1/s_i^2$.

The choice of the number of experiments that are to receive equal weight is to some extent arbitrary. A good working rule is to give equal weight to between 1/2 and 2/3 of the experiments (Cochran, 1937). The method prevents an experiment which happens to have a small estimated error from dominating the result, while allowing the less precise experiments to receive lower weights.

Example 3.

In studies by the U. S. Public Health Service of observers' abilities to count the number of flies which settle momentarily on a grill, each of 7 observers was shown, for a brief period, grills with known numbers of flies impaled on them and asked to estimate the numbers. For a given grill, each observer made 5 independent estimates. The data in table 9 are for a grill which actually contained 161 flies. Estimated variances are based on 4 degrees of freedom each.

The value of Bartlett's χ^2 was 19.9, with 6 degrees of freedom and a significance probability of less than 0.01. Evidently the observers differ in precision. The F-value in the test for interactions was practically 1, giving no indication of any differential bias in observers' error.

The only point of interest in estimating the overall mean is to test whether there is any consistent bias among observers in estimating the 161 flies on the grill. Although inspection of table 9 suggests no such bias, the data will serve to illustrate the application of partial weighting.

It is clear from table 9 that if weighting inversely as s_i^2 were employed,

TABLE 9
OBSERVERS' MEAN ESTIMATES AND ERROR VARIANCES

Observer	Mean estimate x_i	s_i^2	Partial weights
1	183.2	117.0	.0129
2	149.0	8.1	.0129
3	154.0	235.9	.0042
4	167.2	295.0	.0034
5	187.2	1064.6	.0009
6	158.0	51.2	.0129
7	143.0	134.0	.0129
			$w = .0601$

observer 2 would have great influence on the estimate. For partial weighting, we give the same weight to observers 1, 2, 6, and 7. Since $\bar{s}_p^2$ is 77.6 for these observers, $\bar{w}_p = .0129$. The partial weights appear at the right of table 9.

$$\bar{x}_{pw} = \frac{(.0129)(183.2) + \cdots + (.0129)(143.0)}{.0601} = 158.9$$

For the standard error, let

u = no. of experiments given individual weights = 3

w_u = total weight for these u experiments = 0.0085

p = no. of experiments given the same weight = 4

$\bar{n}_p$ = average no. of d.f. for these p experiments = 4

If $\bar{n}_p$ is less than 8,

$$\text{s.e.}(\bar{x}_{pw}) = \frac{1}{w}\sqrt{p\bar{w}_p + \lambda w_u} \tag{19}$$

where λ is read as a function of $\bar{n}_p$ and u from table 12 as described in section 8. In this example, with $\bar{n}_p = 4$ and $u = 3$, $\lambda = 1.8$.

$$\text{s.e.}(\bar{x}_{pw}) = \frac{\sqrt{(4)(.0129) + (1.8)(.0085)}}{.0601} = 4.3$$

For $\bar{n}_p$ greater than 8, the standard error may be taken as approximately

$$\text{s.e.}(\bar{x}_{pw}) = \frac{1}{w}\sqrt{p\bar{w}_p + w_u\left\{1 + \frac{4}{w_u^2}\sum\frac{1}{n_i'}\,w_i(w_u - w_i)\right\}} \tag{20}$$

where the $\sum$ is taken over the u experiments only and

$$n_i' = n_i - \frac{4(u-2)}{(u-1)}$$

Formulas (19) and (20) are revisions of an earlier formula, given by Yates and Cochran (1938), which assumed p and u to be large. In this example, formula (20), although outside of the range of its applicability, agrees well with (19), giving a value of 4.5 for the standard error.

Interactions present.

In this case we again have the model

$$x_i = \mu + (\mu_i - \mu) + e_i$$

and the variance of x_i is $(\sigma_\mu^2 + \sigma_i^2)$. The choice of estimate lies between the unweighted mean and the sample semi-weighted mean $\sum \hat{W}_i x_i / \sum \hat{W}_i$, where

$$\hat{W}_i = \frac{1}{s_\mu^2 + s_i^2}$$

The quantity s_μ^2 is computed by formula

$$s_\mu^2 = \frac{\sum (x_i - \bar{x})^2}{(k-1)} - \bar{s}^2$$

where $\bar{s}^2$ is the mean of the s_i^2.

As explained in section 4, the relative precision of $\bar{x}$ and $\bar{x}_{sw}$ depends on the size of σ_μ^2 and on the amount of variation among the σ_i^2. Use of $\bar{x}$ when F exceeds 4 is a safe working rule, unless there are extremely large variations in the precisions of the individual x_i.

Example 4.

Example 1, previously discussed, represents a situation where the data do not indicate very clearly what kind of model and analysis are appropriate. The probability value for Bartlett's χ^2, 0.06, made it doubtful whether equal precision per observation could be postulated. Although this assumption was adopted in the original analysis, tests for interactions without making this assumption were carried out in section 5. The F and F_w values gave probabilities of 0.09 and 0.07, raising the further question whether interactions should be considered as present or absent. Since, however, the value of F was 2.75, the more

cautious procedure is to recognize that interactions may be present and use a semi-weighted mean. The subsidiary computations needed are given in table 10.

TABLE 10
COMPUTATIONS FOR THE SEMI-WEIGHTED MEAN

Expt.	x_i	s_i^2	$s_\mu^2 + s_i^2$	$10^6 W_i$
1	86.4	1,719	5,142	194
2	158.0	3,348	6,771	148
3	11.3	807	4,230	236
Totals	255.7	5,874		578

$$s_\mu^2 = 5{,}381 - 1{,}958 = 3{,}423$$

The calculation proceeds in the right hand columns of table 10. We find

$$\bar{x}_{sw} = 74.1 : \text{s.e.}(\bar{x}_{sw}) = \frac{10^3}{\sqrt{578}} = 41.6$$

The estimate originally made in example 1 was 85.2 ± 42.3. The two estimates do not agree very closely. The difference is due to an apparently fortuitous correlation between x_i and s_i^2.

The remaining sections deal with the derivation of some of the appropriate formulas.

6. COMPARISON OF THE UNWEIGHTED AND SEMI-WEIGHTED MEANS

Given that interactions are present, the variance of the unweighted mean is

$$V(\bar{x}) = \frac{\sigma_\mu^2 + \bar{\sigma}^2}{k} : \bar{\sigma}^2 = \sum \sigma_i^2/k \qquad (21)$$

If the semi-weights are known exactly, the variance of the semi-weighted mean is

$$V(\bar{x}_{sw}) = \frac{1}{\sum W_i}, \qquad \text{where} \qquad W_i = \frac{1}{\sigma_\mu^2 + \sigma_i^2} \qquad (22)$$

Owing to errors in the weights, the variance of the *sample* semi-weighted mean will be greater than (22). Hence the ratio of (21) to (22) gives an upper limit to the relative precision of $\bar{x}_{sw}$ to $\bar{x}$. This

ratio is

$$\frac{V(\bar{x})}{V(\bar{x}_{sw})} = \frac{(\sigma_\mu^2 + \bar{\sigma}^2)}{k} \sum \left\{ \frac{1}{\sigma_\mu^2 + \sigma_i^2} \right\} = \lambda \qquad \text{(say)} \tag{23}$$

If

$$I = \sigma_\mu^2 / \bar{\sigma}^2$$

is the ratio of the interaction variance to the average error variance, (23) may be written

$$\lambda = \frac{(I+1)}{k} \sum \left\{ \frac{1}{I + \frac{\sigma_i^2}{\bar{\sigma}^2}} \right\} \tag{24}$$

In the development of a working rule about the choice between $\bar{x}$ and $\bar{x}_{sw}$, the first step is to find an upper limit to λ when we fix the two quantities I and the ratio r of the greatest to the smallest error variances. For the following argument, I am indebted to Dr. Paul Meier.

Let σ_1^2 be the smallest error variance, and let

$$\sigma_i^2 = r_i \sigma_1^2 \qquad (1 \leq r_i \leq r; i > 1)$$

Then

$$k\bar{\sigma}^2 = \sigma_1^2(r_1 + r_2 + \cdots + r_k) = \sigma_1^2 R \qquad \text{(say)}$$

Hence (24) becomes

$$\lambda = \frac{(I+1)}{k} \sum \left\{ \frac{1}{I + \frac{kr_i}{R}} \right\} \tag{25}$$

$$= \frac{(I+1)}{k^2} \sum \left\{ \frac{R}{(AR + r_i)} \right\} \tag{26}$$

where $A = I/k$.

The argument proceeds by showing that λ cannot have a maximum unless every r_i is at one of the ends of its possible range from 1 to r. Since I is fixed, we may neglect the term $(I+1)/k^2$ and consider the quantity

$$\gamma = \sum \left\{ \frac{R}{AR + r_i} \right\} \tag{27}$$

It may be verified that

$$\frac{\partial \gamma}{\partial r_h} = \sum{}' \left\{ \frac{r_i}{(AR + r_i)^2} \right\} - \frac{R'}{(AR + r_h)^2}$$

$$\frac{\partial^2 \gamma}{\partial r_h^2} = \frac{2R'(A + 1)}{(AR + r_h)^3} - 2A \sum{}' \left\{ \frac{r_i}{(AR + r_i)^3} \right\}$$

where a prime denotes summation over all terms except that in r_h .

Hence, at any point at which $\partial\gamma/\partial r_h = 0$, we have

$$\frac{\partial^2 \gamma}{\partial r_h^2} = \frac{2(A + 1)}{(AR + r_h)} \sum{}' \left\{ \frac{r_i}{(AR + r_i)^2} \right\} - 2A \sum{}' \left\{ \frac{r_i}{(AR + r_i)^3} \right\}$$

$$= \frac{2(A + 1)}{(AR + r_h)} \sum{}' \frac{r_i}{(AR + r_i)^2} \left\{ 1 - \frac{A}{(A + 1)} \frac{(AR + r_h)}{(AR + r_i)} \right\}$$

The term inside the curly brackets is easily seen to be positive for every i. Hence at any point where $\partial\gamma/\partial r_h = 0$, we have the second derivative positive, so that there is no interior maximum.

To find the maximum value of λ, let m of the r_i be 1, and the remaining $(k - m)$ be r. Then from (26)

$$\lambda = \frac{R(I + 1)}{k^2} \left\{ \frac{m}{AR + 1} + \frac{k - m}{AR + r} \right\}$$

where now

$$R = \sum r_i = m + (k - m)r$$

There is no convenient analytic expression for the maximizing value of m, but for given I and r the maximum is easily computed numerically. The results in table 11 show the reciprocal of the maximum, i.e. the lower bound to the relative precision of $\bar{x}$ to $\bar{x}_{sw}$.

TABLE 11

LOWER LIMITS OF RELATIVE PRECISION OF $\bar{x}$ TO $\bar{x}_{sw}$

r = largest/ smallest error variance	I = ratio of interaction variance to average error variance			
	0	1	2	3
2	.89	.97	.99	.99
3	.75	.93	.97	.98
4	.64	.90	.95	.97
6	.49	.85	.93	.95
8	.40	.81	.90	.94
16	.22	.74	.86	.91

If an upper limit of 10 per cent in the loss of precision is regarded as tolerable, table 11 shows that the unweighted mean is satisfactory whenever I exceeds 3, or when I is at least 2 and r is 8 or less, or when I is at least 1 and r is 4 or less. Values of r greater than 16 were not included in the table, on the grounds that such cases would represent a very extreme degree of variation in the σ_i^2 .

The translation of these results into the working rules given in sections 4 and 5 can be made only approximately, since in practice we do not know the value of I. In section 5, the numerator of F is an unbiased estimate of $(\sigma_\mu^2 + \bar{\sigma}^2)$, while the denominator is an unbiased estimate of $\bar{\sigma}^2$. Hence, F can be considered as an estimate of $(1 + I)$, although the estimate may be shown to be positively biased. The rule given in section 5, namely to use $\bar{x}$ in general when F exceeds 4, was chosen because with $F > 4$, I is unlikely to be < 2, and from table 11 the unweighted mean suffers little loss for $I = 2$ unless the x_i differ greatly in precision.

Similarly, the F-ratio in table 3 of section 4 is an estimate of

$$1 + \frac{I\bar{f}'}{k} \sum \left(\frac{1}{f_i}\right)$$

If the range of values of f_i is not too great, this expression is approximately $(1 + I)$, and leads to the rules given in table 4.

These rules are perhaps biased in favor of the semi-weighted mean, because the figures in table 11 are underestimates of the relative precision of $\bar{x}$ to $\bar{x}_{sw}$.

7. APPROXIMATE NUMBER OF DEGREES OF FREEDOM IN THE STANDARD ERROR OF $\bar{x}$

In sections 4 and 5 the formula

$$\frac{s_b}{\sqrt{k}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{k(k-1)}}$$

has been recommended for the standard error of $\bar{x}$ when interactions are present. Since the x_i do not have equal variances, this standard error is not distributed in the usual way for a root mean square with $(k - 1)$ degrees of freedom.

The distribution of s_b^2 will, however, be approximated by a distribution of the usual type for a mean square. The number of degrees of freedom n_e ascribed to this distribution will be chosen so as to give it the correct variance.

Let

$$\theta_i = \sigma_\mu^2 + \sigma_i^2$$

$$\Theta_1 = \frac{\sum \theta_i}{k} \quad : \quad \Theta_2 = \frac{\sum \theta_i^2}{k}$$

If the x_i are normally and independently distributed about μ, the variance of s_b^2 is found by algebra to be

$$V(s_b^2) = 2\{(k-2)\Theta_2 + \Theta_1^2\}/(k-1)^2 \tag{28}$$

For the typical distribution of a mean square (i.e. that of a multiple of χ^2 with n_e degrees of freedom),

$$V(s_b^2) = \frac{2\{E(s_b^2)\}^2}{n_e} = \frac{2\Theta_1^2}{n_e} \tag{29}$$

Hence, by equating (28) to (29),

$$n_e = \frac{(k-1)^2\Theta_1^2}{(k-2)\Theta_2 + \Theta_1^2} \tag{30}$$

In practice, we must substitute the sample estimates of Θ_1 and Θ_2 .

8. THE STANDARD ERROR OF THE WEIGHTED MEAN WHEN THE n_i ARE SMALL

Meier's formula (13) or (14) in section 5 is satisfactory for values of n_i down to 8 or down to 6 when k is small. For very small values of n_i , the value of the factor in curly brackets in (13) and (14) has been estimated by experimental sampling, Cochran and Carroll, (1953). Values taken from this paper appear in table 12.

TABLE 12

VALUES OF λ FOR WHICH λ/w IS AN ESTIMATE OF $V(\bar{x}_w)$

	Number of Experiments									
$\bar{n}$	2	3	4	5	6	8	10	12	15	20
2	2.0	2.9	3.9	5.1	6.1	7.9	10.6	12.6	17.1	22.8
4	1.5	1.8	2.2	2.5	2.7	3.2	3.7	4.1	4.7	5.4
6	1.3	1.5	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4
8	1.2	1.5	1.5	1.6	1.6	1.7	1.8	1.8	1.9	1.9

The use of this formula is illustrated in section 9.

9. STANDARD ERROR OF THE PARTIALLY WEIGHTED MEAN

Let $\bar{x}_p$ be the mean of the p experiments which each receive equal weight $\bar{w}_p$, and $\bar{x}_u$ be the weighted mean of the remaining u experiments which receive individual weights. For the estimated variance of $\bar{x}_p$ we take the average observed variance in these experiments, divided by p; or in other words,

$$s^2(\bar{x}_p) = \frac{1}{p\bar{w}_p}$$

For the estimated variance of $\bar{x}_u$, we use either Meier's formula or, if the n_i are less than 8, the empirical formula in the previous section,

$$s^2(\bar{x}_u) = \frac{\lambda}{w_u}$$

where λ is read from table 12 and w_u is the total weight for these u experiments.

The overall mean is

$$\bar{x}_{pw} = \frac{p\bar{w}_p\bar{x}_p + w_u\bar{x}_u}{w}$$

Hence, if $\bar{w}_p$ and w_u can be regarded as free from error,

$$s(\bar{x}_{pw}) = \frac{\sqrt{(p\bar{w}_p)^2 s^2(\bar{x}_p) + w_u^2 s^2(\bar{x}_u)}}{w}$$

$$= \frac{\sqrt{p\bar{w}_p + \lambda w_u}}{w}$$

as given in section 5.

This argument is non-rigorous in several ways. The variance given for $\bar{x}_p$ is too low, since these p experiments are selected because they appear to be precise. Similarly, the variance for $\bar{x}_u$ is too high. Also, errors in the relative weights, $p\bar{w}_p$ and w_u, are ignored. However, the formula does reduce to the appropriate values when $p = k$ and when $p = 0$: in intermediate cases my guess is that it may be slightly too low.

10. SUMMARY

This paper discusses methods for combining a number of estimates x_i of some quantity μ, made in different experiments. For the ith estimate we have an unbiased estimate s_i^2 of its variance, based on n_i degrees of freedom.

It is important to find out whether the x_i agree with one another

within the limits of their experimental errors. If they do not, i.e. if interactions are present, the type of overall mean that will be useful for future action requires careful consideration. However, in most cases the problem of estimating the mean of the x_i , at least over some subgroup of the experiments, will remain.

If the experiments are of the same type and the x_i are of equal precision, the best estimate in general is the unweighted mean $\bar{x}$, but its standard error differs according as interactions are present or absent.

The second case considered is that in which the experiments are of different types, but the variance σ_0^2 per observation is the same in all experiments. The variance of x_i is then of the form σ_0^2/f_i . If there are no interactions, the best combined estimate is the weighted mean $\sum f_i x_i / \sum f_i$. If interactions exist, the choice lies between the unweighted mean and a semi-weighted mean. Recommendations for this choice are given. In the semi-weighted mean, the weights W_i are ideally inversely proportional to

$$\sigma_\mu^2 + \frac{\sigma_0^2}{f_i}$$

The semi-weighted mean reduces to the weighted mean when the interaction variance $\sigma_\mu^2 = 0$, and to the unweighted mean when the interaction variance is large. In practice, sample estimates of the weights are used.

Experiments in which the variance per observation is not constant represent perhaps the most common case in practice. In the absence of interactions, possible estimates are the unweighted mean, weighting inversely as s_i^2 or, if the n_i are small, a kind of partial weighting. When interactions are present, the unweighted mean or the semi-weighted mean is appropriate. Working rules are given to aid in the selection of an estimate.

In conclusion, the unweighted mean will probably be satisfactory with many sets of data. The principal value in learning about various types of more complex estimates lies in occasional situations in which the unweighted mean would incur a substantial loss of precision, and also in receiving assurance that the unweighted mean is often entirely adequate.

Some approximations in theory that are needed for the practical recommendations are developed in later sections of the paper.

REFERENCES

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Stat.*, *15*, 190–204.

Bliss, C. I. (1952). *The statistics of bioassay.* Academic Press Inc., New York. p. 576.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Jour. Roy. Stat. Soc., Supp., 4*, 102–118.

Cochran, W. G. and Carroll, S. P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics, 9*, 447–459.

Cochran, W. G. and Cox, G. M. (1950). *Experimental designs*. John Wiley and Sons, Inc., New York. Chapter 14.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38*, 324–329.

Kempthorne, O. (1952). *The design and analysis of experiments*. John Wiley and Sons, Inc., New York. Chapter 28.

Meier, P. (1953). Variance of a weighted mean. *Biometrics, 9*, 59–73.

Paull, A. E. (1950). On a preliminary test for pooling mean squares in the analysis of variance. *Ann. Math. Stat., 21*, 539–556.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biom. Bull., 2*, 110.

Welch, B. L. (1951). On the comparison of several mean values—an alternative approach. *Biometrika, 38*, 330–336.

Whitlock, J. H. and Bliss, C. I. (1943). A bioassay technique for anthelmintics. *Jour. Parisitology, 29*, 48–58.

Yates, F. and Cochran, W. G. (1938). The analysis of groups of experiments. *Jour. Agr. Sci., 28*, 556–580.