CHAPTER 4

Further Aspects of Planning

4.1 SAMPLE SIZE IN RELATION TO TESTS OF SIGNIFICANCE

Suppose that the investigator has identified two or more groups of subjects whose mean values for some response variable y he wishes to compare. A decision must be made about the size of sample to be selected from each group. Sometimes this decision is controlled largely by cost or availability considerations. A group that is of particular interest may contain only 80 subjects, or the budget may limit the study to two samples of sizes not exceeding 200 each. In the absence of such limitations, statistical theory provides certain formulas as a guide in making decisions about sample size. Use of these formulas may present difficulties, either because the formulas oversimplify the actual conditions of the survey or because the investigator does not have certain information about the study that the formulas require. Nevertheless, it is worth finding out what light these formulas throw on the sample-size issue even when the size is limited by costs or availability.

Calculation of sample size in relation to a test of significance is most often made in exploratory studies. Suppose that there are two groups of subjects exposed to different agents, or one group exposed to an agent and a control group unexposed. If the study fails to find a significant difference $\bar{y}_1 - \bar{y}_2$, the investigator knows that he will have obtained an inconclusive result. The group means have not been shown to be different, but neither have they been shown to be essentially the same, since this conclusion would amount to assuming that the null hypothesis has been proved correct, or nearly correct.

If $\bar{d} = \bar{y}_1 - \bar{y}_2$ and δ is the unknown population difference between the group means, the investigator might reason that he does not mind finding $\bar{d}$

nonsignificant if δ is small. However, if δ is large enough to be of practical importance, the investigator wants to have a high probability of detecting that there is a difference by finding $\bar{d}$ significant. Such a result may encourage later work that will estimate δ more accurately. This leads to the question: For a given δ and given sample sizes from the two groups, what is the probability of obtaining a significant $\bar{d}$? This probability is called *the power of the test.*

This calculation is easily made if bias can be ignored and if we can assume $\bar{d}$ normally distributed with mean δ and standard deviation $\delta_{\bar{d}}$. Let z_α be a nonnegative number such that the probability that a normal deviate exceeds z_α is α. For example, if $z_\alpha = 0$, $\alpha = 0.5$ and if $z_\alpha = 1.96$, then $\alpha = 0.025$, since we are considering only the right-hand tail of the normal distribution.

We can now calculate the probability that $\bar{d}$ is significant. We start with a one-tailed test (δ assumed $\geqslant 0$), since this is slightly easier. Clearly, $\bar{d}$ is significant if it exceeds $z_\alpha\sigma_{\bar{d}}$. Thus, we want to find the probability that $\bar{d}$ exceeds $z_\alpha\sigma_{\bar{d}}$. If $\bar{d}$ is normally distributed with mean δ and standard deviation $\sigma_{\bar{d}}$, the standard normal deviate corresponding to $\bar{d}$ is therefore $(\bar{d} - \delta)/\sigma_{\bar{d}}$. Now setting $\bar{d} = z_\alpha\sigma_{\bar{d}}$, the threshold significant value, we compute the probability that $\bar{d}$ is significant by calculating

$$z = \frac{z_\alpha\sigma_{\bar{d}} - \delta}{\sigma_{\bar{d}}} = z_\alpha - \frac{\delta}{\sigma_{\bar{d}}} \qquad (4.1.1)$$

and reading the probability that a normal deviate exceeds z. If we do this for fixed n and α and various δ, we produce a curve called *the power function* which relates power and δ.

If the study is planned to have two independent groups, each of n subjects, then $\sigma_{\bar{d}} = \sqrt{2}\,\sigma/\sqrt{n}$, where σ is the standard deviation per subject in both groups. Formula (4.1.1) then becomes

$$z = z_\alpha - \sqrt{\frac{n}{2}}\,\frac{\delta}{\sigma} \qquad (4.1.2)$$

Let us consider some examples which illustrate the use of formulas in the estimation of sample size.

Example 1. Suppose that costs limit the sample sizes to $n = 100$. Related data indicate that σ is about 1. The investigator thinks that if δ is as large as 0.3, this is important enough so that he would like to obtain a significant difference. What is the probability? In this case $z_\alpha = 1.64$ for a one-tailed test at the 5% level. Thus by (4.1.2),

$$z = 1.64 - \sqrt{50}\,(0.3) = 1.64 - 2.12 = -0.48$$

The normal tables give $P = 0.68$ for the probability of exceeding z, a little disappointing, but as good as many studies can offer.

In a two-tailed test, $\bar{d}$ can be either significantly positive or significantly negative. In our notation the conditions are $\bar{d} > z_{\alpha/2}\sigma_{\bar{d}}$ or $\bar{d} < -z_{\alpha/2}\sigma_{\bar{d}}$. Note the subscript $\alpha/2$; if the two-tailed probability is to be 0.05, the one-tailed probability must be 0.025. If $\delta > 0$, a verdict that $\bar{d} < -z_{\alpha/2}\sigma_{\bar{d}}$ would be a horrible mistake, since we find $\bar{d}$ significant, but in the wrong direction. Fortunately, if the probability that $\bar{d}$ is significant in the correct direction is at all sizable (e.g., > 0.2), the probability that $\bar{d}$ is significant in the wrong direction is tiny and can be ignored. Hence in a two-tailed test we can calculate the probability that $\bar{d}$ is significant and in the correct direction by amending (4.1.1) to the probability that a normal deviate exceeds

$$z = z_{\alpha/2} - \frac{\delta}{\sigma_{\bar{d}}} \tag{4.1.3}$$

With two planned samples each of size n, we can now calculate the needed value of n such that the probability of finding a significant $\bar{d}$ has any desired value β. Take $\beta > 0.5$, since we want the probability to be high. Earlier, we defined z_α ($\geqslant 0$) as a value for which the probability that a normal deviate exceeds z_α is α. This definition restricts us to values of $\alpha \leqslant 0.5$. If $\beta > 0.5$, the value of z such that the probability β of exceeding this z is $-z_{(1-\beta)}$. By the symmetry of the normal curve, the probability that $z \leqslant -z_{(1-\beta)}$ is $(1-\beta)$, so that the probability that $z > -z_{(1-\beta)}$ is β for $\beta > 0.5$.

To summarize, if we want a *one-tailed* test to have probability β (> 0.5) of finding a significant result at level α, we write

$$-z_{1-\beta} = z_\alpha - \sqrt{\frac{n}{2}}\,\frac{\delta}{\sigma} \tag{4.1.4}$$

and solve for n, giving

$$n = \frac{2(z_\alpha + z_{1-\beta})^2\sigma^2}{\delta^2} \tag{4.1.5}$$

If the test is *two-tailed*,

$$n = \frac{2(z_{\alpha/2} + z_{1-\beta})^2\sigma^2}{\delta^2} \tag{4.1.6}$$

For one- and two-tailed tests, Table 4.1.1 shows the multipliers of σ^2/δ^2 for specified probabilities, from 0.5 to 0.95, of finding a significant difference.

As before, the ratio δ/σ that is of importance must be specified in order to use Table 4.1.1.

Example 2. A pilot study suggests that σ may be about 6 while the investigator would like $\beta = 0.95$ if δ is 2 in a two-tailed test at the 5% level. For this, $\sigma^2/\delta^2 = 9$ and $n = (26.6)(9) = 239$ in each sample.

Table 4.1.1 may also be used as an approximation when the response is a binomial proportion and we are comparing independent samples from two populations whose proportions are p_1 and p_2. The numerical factors in Table 4.1.1 remain the same, but σ^2/δ^2 is replaced by

$$\frac{p_1q_1 + p_2q_2}{2(p_1 - p_2)^2} \tag{4.1.7}$$

with $q = 1 - p$, or $q = 100 - p$ if p is expressed as a percentage. If the first population is a control or a standard method, p_1 may be known fairly well from previous studies. The value of p_2 has to be inserted from consideration of the size of difference $|p_1 - p_2|$ that the investigator does not want to "miss" in the sense of this test of significance.

Example 3. If $p_1 = 6\%$, $p_2 = 3\%$, then $q_1 = 94\%$, $q_2 = 97\%$. To have an 80% chance of finding a significant difference in a one-tailed test, we require

Table 4.1.1. Multipliers of σ^2/δ^2 Needed to Give n for a Specified Probability of Finding a 5% Significant Difference in the Correct Direction

Probability	One-Tailed Test	Two-Tailed Test
0.5	5.4	8.0
0.6	7.2	10.2
0.7	9.4	12.7
0.8	12.4	16.2
0.9	17.1	21.5
0.95	21.6	26.6

the size of each sample to be

$$n = \frac{(12.4)[(6)(94) + (3)(97)]}{(2)(9)} = 589$$

In this and the preceding section the sample-size formulas that involve n assume two *independent* samples. Often, the samples are matched or paired by certain characteristics of the subjects. If so, the quantity σ^2/δ^2 in Table 4.1.1 is replaced by $\sigma^2(1 - \rho)/\delta^2$, where ρ is the correlation coefficient between members of the same pair. Sometimes ρ can be guessed if an estimate of σ^2 is available. Alternatively, if $d_j = y_{1j} - y_{2j}$, the difference between the members of the jth pair, σ^2 may be replaced by $\sigma_d^2/2$. If σ_d^2 is being estimated from a past study, this should of course have employed the same criteria for matching. With binomial data, a fair amount of evidence suggests that pairing is usually only moderately effective in increasing precision. Calculation of n by formula (4.1.6) for independent samples will be on the conservative side, but not badly wrong.

The method extends to cases not so simple as two samples of size n, provided that $\sigma_{\bar{d}}$ can be calculated. We give two numerical illustrations in Example 4 and a general algebraic one in Example 5.

Example 4. The group on which a treatment acts will provide only 50 subjects. The control group is not so restricted. The investigator guesses that the probability of detecting his desired δ will not be high if only 50 control subjects are used. How much better does the investigator do if n is 100 or 200 for the control sample? We have $\sigma = 10$, $\delta = 4$. We revert to formula (4.1.3), with $z_{\alpha/2} = 1.96$ in place of z_α since a two-tailed test is desired.

$$z = 1.96 - \frac{4}{\sigma_{\bar{d}}}$$

With 50 control subjects, $\sigma_{\bar{d}} = \sqrt{2}\,\sigma/\sqrt{50} = 2$, so that $z = -0.04$, giving a probability 0.52. With 100 and 200 controls,

$$\sigma_{\bar{d}} = \sigma\sqrt{\frac{1}{50} + \frac{1}{100}} = 1.732; \qquad \sigma_{\bar{d}} = \sigma\sqrt{\frac{1}{50} + \frac{1}{200}} = 1.581$$

so that $z = -0.35$ and -0.57, with probabilities 0.64 and 0.72, respectively.

Example 5. In what is called a before–after study, measurements are taken in each group both before an agent has been applied to the group and at some time afterwards. The quantity of interest is often

$$\bar{d} = (\bar{y}_{1a} - \bar{y}_{1b}) - (\bar{y}_{2a} - \bar{y}_{2b})$$

the mean difference in the changes associated with each agent. If each measurement has the same variance σ^2, and the correlation between before and after measurements in the same group is ρ, then

$$\bar{y}_{1a} - \bar{y}_{1b} = \bar{d}_1 \quad \text{and} \quad \bar{y}_{2a} - \bar{y}_{2b} = \bar{d}_2$$

$$\text{Var}\,\bar{d}_1 = \frac{\sigma^2 - 2\rho\sigma^2 + \sigma^2}{n} = \frac{2\sigma^2}{n}(1 - \rho) = \text{Var}\,\bar{d}_2$$

Combining this information, we find

$$\sigma_{\bar{d}} = \sqrt{\sigma_{\bar{d}_1}^2 + \sigma_{\bar{d}_2}^2} = \sigma\sqrt{\frac{4}{n}(1 - \rho)}$$

This $\sigma_{\bar{d}}$ is used in formula (4.1.1).

4.2 SAMPLE SIZE FOR ESTIMATION

Sometimes investigators prefer to look at sample-size formulas from the viewpoint of closeness of estimation rather than of testing significance. This is so, for instance, if there is a good deal of presumptive evidence in advance that a treatment will produce some effect; the question is whether our estimates will be sufficiently accurate as a basis for action. As before, we assume that bias is negligible and that our estimated difference $\bar{d}$ can be taken to be normally distributed. Thus, if δ is the population difference, $\bar{d}$ should lie within the limits $\delta \pm L$ with about 95% probability where

$$L = 2\sigma_{\bar{d}} \tag{4.2.1}$$

In the simplest application to two independent samples each of size n

$$L = \frac{2\sqrt{2}\,\sigma}{\sqrt{n}} = \frac{2.82\sigma}{\sqrt{n}} \tag{4.2.2}$$

where σ is the within-group standard deviation. Formulas (4.2.1) and (4.2.2) can be used in a number of ways.

Let us consider examples which illustrate the use of formulas to produce the sample-size estimate correct to a certain limit with high probability.

Example 1. If $\bar{d}$ is desired to be correct to within specified limits of error $\pm L$ (apart from a 1 in 20 chance), we have from (4.2.2)

$$n = \frac{8\sigma^2}{L^2} \qquad (4.2.3)$$

as the size of each group.

Example 2. Suppose that financial or other considerations limit n to 400 and that σ is thought to be about 3. From (4.2.2)

$$L = \frac{2.82\sigma}{\sqrt{n}} = 0.423$$

Consideration as to whether this is satisfactory will probably involve some thought about any action to be taken. The situation might be "I feel that action is necessary if $\delta \geqslant 1$ and will argue for action if $\bar{d} \geqslant 1$, but not otherwise." Clearly, if we are unlucky we may find ourselves arguing for action if δ is only $(1 - 0.423) = 0.577$, or failing to argue for action when δ is nearly as high as 1.423. The issue then depends on whether mistakes of this kind are regarded as tolerable.

Example 3. If the response variable is a binomial proportion so that $\bar{d} = \hat{p}_1 - \hat{p}_2$, then with independent samples, $\sqrt{2}\,\sigma$ becomes $\sqrt{p_1q_1 + p_2q_2}$. Hence (4.2.2) becomes

$$L = \frac{2\sqrt{p_1q_1 + p_2q_2}}{\sqrt{n}} \qquad (4.2.4)$$

This formula holds whether p_1, p_2, and L are all expressed in proportions or percentages. In proportions, $q_i = 1 - p_i$; in percentages, $q_i = 100 - p_i$.

Suppose samples of $n = 3600$ can be run and the failure rate (response) in the control group is 10%. The failure rate in the treatment group is not known, but if it is as low as 5%, how well is the improvement in failure rate estimated? With $p_1 = 10, p_2 = 5$,

$$L = \tfrac{2}{60}\sqrt{(10)(90) + (5)(95)} = 1.24\%$$

This might seem good enough. Even if the treatment is ineffective, $p_2 = 10\%$, we have

$$L = \tfrac{2}{60}\sqrt{(10)(90) + (10)(90)} = 1.41\%$$

so that it is unlikely that p_2 would be estimated as more than 1.41% lower than p_1.

Example 4. When estimating a treatment effect from two groups of subjects, the investigator may have some subgroups for which it would be informative (1) to estimate the treatment effect separately in each subgroup, and (2) to compare the sizes of the treatment effects for different subgroups. Unfortunately, as is well known, much larger sample sizes are needed for case (1) and particularly for case (2) than for the estimation of an overall treatment effect.

Suppose for illustration two subgroups contain proportions ϕ and $1 - \phi$ of the subjects in each of a treatment and a control population. With male–female subgroups, ϕ might be about 0.5; with white–black subgroups, ϕ might be 0.8 or 0.9. The values of $\sigma_{\bar{d}}$ are therefore $\sqrt{2}\,\sigma/\sqrt{n}$ for the overall effect, approximately $\sqrt{2}\,\sigma/\sqrt{\phi n}$ and $\sqrt{2}\,\sigma/\sqrt{(1-\phi)n}$ for the individual subgroup effects, and $\sqrt{2}\,\sigma/\sqrt{\phi(1-\phi)n}$ for the difference in effect from one subgroup to the other. Since $L = 2\sigma_{\bar{d}}$, the multipliers of σ^2/L^2 required to find n are 8, $8/\phi$, $8/(1-\phi)$, and $8/\phi(1-\phi)$. These multipliers show the relative sample sizes needed to attain the same error limit $\pm L$, and are presented in the list below for $\phi = 0.5$ and 0.9.

ϕ	Overall Effect	Effect in Subgroup 1	Effect in Subgroup 2	Difference in Effects
0.5	8	16	16	32
0.9	8	9	80	89

With subgroups of equal size, the most favorable case, estimation of effects separately in each subgroup requires twice the sample size, while estimating the difference in effects requires four times the size. With k equal subgroups the multipliers are k and $2k$. As the case $\phi = 0.9$ illustrates, the situation is much worse when some subgroups are relatively small.

These results are not intended to deter an investigator from examining effects separately in different subgroups. But the accompanying standards of precision are lower, and large samples are usually needed to estimate a difference in effects from one subgroup to another.

Example 5. If the cost of sampling and measurement is much cheaper in population 1 than in population 2, the question is occasionally asked: What sample sizes n_1 and n_2 will provide a specified value of $V(\bar{d})$ at minimum cost? Let

$$\text{cost} = C = c_1n_1 + c_2n_2 \ (c_1 < c_2); \qquad V = V(\bar{d}\,) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The calculus minimum-cost solution is

$$\frac{n_1}{\sigma_1\sqrt{c_2}} = \frac{n_2}{\sigma_2\sqrt{c_1}} = \frac{n}{\sigma_1\sqrt{c_2} + \sigma_2\sqrt{c_1}}; \qquad n = \frac{\left(\sigma_1\sqrt{c_2} + \sigma_2\sqrt{c_1}\right)^2}{V\sqrt{c_1c_2}}$$

Assuming σ_1 and σ_2 are roughly equal, we have $n_1/n_2 = \sqrt{c_2}/\sqrt{c_1}$. Unless the cost-ratio c_2/c_1 is extreme, however, the saving over equal sample sizes is modest, since

$$\frac{C_{\text{equal}}}{C_{\text{min}}} = \frac{2(c_1 + c_2)}{\left(\sqrt{c_1} + \sqrt{c_2}\right)^2}$$

$$= \frac{2(1 + c_2/c_1)}{\left(1 + \sqrt{c_2/c_1}\right)^2}$$

Equal sample sizes cost only 3% more if $c_2/c_1 = 2$, 11% more if $c_2/c_1 = 4$, and 27% more if $c_2/c_1 = 10$.

4.3 THE EFFECT OF BIAS

As mentioned, the formulas in the preceding sections assume that any bias in the estimates is negligible. The effect of bias on the accuracy of estimation was discussed in Section 2.5. To cite results given there, suppose n has been determined so that the probability is 0.95 that $\bar{d}$ lies in the interval $(\delta - L, \delta + L)$ in the absence of bias. The presence of an unsuspected bias of amount $\leqslant 0.2L$ decreases the 0.95 probability only trivially. If $B = 0.5L$, the 0.95 probability is reduced to 0.84 and to less than 0.50 if B/L exceeds 1.0. For given $f = B/L$ a table can be constructed which shows the amount n must be increased over n_0 in the "no bias" situation in order to keep this probability at 0.95. Table 4.3.1 shows the ratio n/n_0 for $f = 0.2(0.1)0.9$.

It is not likely that Table 4.3.1 can be used for estimating n in planning a specific survey, because of ignorance of the value of f. If an investigator somehow knew f fairly well, he/she would try to adjust $\bar{d}$ in order to remove the bias and would then face a different estimation problem. The table helps, however, in considering a possible trade-off between reduction of bias and reduction of random sources of error. For instance, if by better planning or more-accurate measurements the value of f could be reduced from 0.6 to 0.3, a sample size of $1.40n_0$ would be as effective as one of $4.22n_0$, but would be only about one-third the size. A more expensive method of data collection may save money if it reduces the bias sufficiently.

Table 4.3.1. Ratio of n/n_0 Needed in Order to Give 95% Probability that $\bar{d}$ is Correct to Within $\delta \pm L$ When Bias of Amount fL is Present. (Note that n_0 is the sample size in the "no bias" case.)

$f = B/L$	n/n_0
0.2	1.13
0.3	1.40
0.4	1.89
0.5	2.72
0.6	4.22
0.7	7.51
0.8	16.9
0.9	67.6

4.4 MORE COMPLEX COMPARISONS

The illustrations of sample-size problems in Sections 4.1 and 4.2 have referred mainly to the difference between the means of two groups. In studies with more than two groups the comparison of primary interest may be more complex. The procedure here is to define $\bar{d}$ as the estimated comparison of interest, with δ as the population value of the comparison. For a specified probability β (> 0.5) of "detecting" δ, we may rewrite formula (4.1.4) more generally as

$$-z_{(1-\beta)} = z_\alpha - \frac{\delta}{\sigma_{\bar{d}}} \tag{4.4.1}$$

If we want $|\bar{d} - \delta| \leqslant L$ apart from a 1 in 20 chance, we can use (4.2.1),

$$L = 2\sigma_{\bar{d}} \tag{4.4.2}$$

From the nature of the comparison we should be able to express $\sigma_{\bar{d}}$ in terms of n, the size of each group, and then solve for n from (4.4.1) or (4.4.2).

The following examples involve studies having more than two groups.

Example 1. A study has three groups, containing amounts 0, 1, and 2 or a, $a + 1$, and $a + 2$ of the treatment. If the response is thought to be linearly related to the amount of the treatment, the quantity of primary interest may be the average change in response per unit increment in amount. For this,

$$\bar{d} = \frac{\bar{y}_3 - \bar{y}_1}{2} \quad \text{and} \quad \sigma_{\bar{d}} = \frac{\sigma}{\sqrt{2n}}$$

With four groups having amounts 0, 1, 2, and 3 or a, $a + 1$, $a + 2$, and $a + 3$ of the treatment, the corresponding estimate of the average change in y per unit increase in amount is

$$\bar{d} = \frac{3\bar{y}_4 + \bar{y}_3 - \bar{y}_2 - 3\bar{y}_1}{2} \quad \text{and} \quad \sigma_{\bar{d}} = \frac{\sigma\sqrt{5}}{\sqrt{n}}$$

More generally, suppose that we have k groups having amounts $x_1, x_2, \ldots, x_k$. We use weights

$$w_i = x_i - \bar{x}$$

and

$$\bar{d} = \sum w_i \bar{y}_i$$

with variance

$$V(\bar{d}) = \frac{\sum w_i^2 \sigma^2}{n} = \frac{\sigma^2}{n} \sum (x_i - \bar{x})^2$$

Example 2. This example is artificial, but illustrates the way in which the formulas are adapted for regression studies. A firm has several factories doing similar work. Certain tasks sometimes done by the workers require a high degree of skill, concentration, and effort, and good performance in these tasks is important. The management finds that one factory offers one unit per hour of extra pay as incentive for this work, another factory offers three units per hour, and a third factory offers no incentive pay.

The management considers taking a random sample of workers in each of these three factories and recording performance scores. It is proposed to estimate the average change in performance per unit extra incentive pay. If the true increase in performance per unit incentive pay is at least 4%, the management would like a 95% chance of declaring the estimated increase to be significant (5% one-tailed test). What sample size is needed in each factory?

Assuming a linear effect of the incentive-pay performance, the values of x are 0, 1, and 3, with $\Sigma(x - \bar{x})^2 = 14/3$. Hence $\sigma_{\bar{d}} = \sigma\sqrt{14/3n}$. From (4.4.1),

$$\sqrt{\frac{3n}{14}}\,\frac{\delta}{\sigma} = z_\alpha + z_{1-\beta}$$

$$n = \frac{14}{3}(z_\alpha + z_{1-\beta})^2 \frac{\sigma^2}{\delta^2}$$

The management wants $\delta = 0.04\mu$, where μ is the performance level under no incentive pay. For a one-tailed 5% test and probability 0.95, z_α and $z_{1-\beta}$ are both 1.64, so that

$$n = \frac{14(3.28)^2}{(3)(0.0016)}\left(\frac{\sigma}{\mu}\right)^2 \approx 31400\left(\frac{\sigma}{\mu}\right)^2.$$

Work performance scores are not kept routinely, but some recent data indicate a between-workers coefficient of variation (100 σ/μ) of 15% for nonincentive workers. Thus $\sigma/\mu = 0.15$, giving $n = 706.5$ for the sample from each of the three factories. This investigation may be more expensive than the manufacturer is willing to pay. If we reduced the power from 0.95 to 0.5, the sum of the z's would reduce to 1.64, which is half of 3.28. This would reduce the sample sizes to $706/4 \approx 176$, a considerable reduction in effort, but at a large price in ability to detect an improvement.

Some sample-size problems require distributions different from the normal; solutions are sometimes available from results in the literature. For instance, an investigator might be primarily interested in comparing the amounts of variability in two groups as estimated by the sample variances. A rough answer to this problem can be obtained from tables of the F distribution, though this assumes normality in the original distribution of y and the sizes have to be increased substantially if the distribution of y is long-tailed, with positive kurtosis.

4.5 SAMPLES OF CLUSTERS

The illustrations in Sections 4.1 and 4.2 may be unrealistic for a second reason, in that the structure of the sample is more complex than has been assumed. A common case is that in which the individuals in a sample fall naturally into subgroups or clusters, the sample being drawn by clusters.

These clusters might be families, Boy Scout troops, Rotary Clubs, schools, or church congregations.

Sampling now proceeds in two stages. First, a sample of k clusters is chosen at random; then from each chosen cluster a sample of individuals is randomly drawn, n_j of them from cluster j. Letting y_{jr} represent the rth observation in the jth sampled cluster, we fix ideas by writing

$$y_{jr} = \mu + \gamma_j + e_{jr} \qquad \left(j = 1, 2, \ldots, k;\ r = 1, 2, \ldots, n_j; \qquad \sum_{j=1}^{k} n_j = n \right) \tag{4.5.1}$$

In Eq. (4.5.1), μ is the population mean. γ_j is the departure from μ of the jth cluster's mean so γ_j is a random quantity with mean zero and a variance that we shall name σ_γ^2, the between-cluster variance. The value e_{jr} is the random departure of y_{jr} from its own cluster mean; thus e_{jr} also has mean zero, and we shall use σ_{ej}^2 for its variance, noting that this within-cluster variance may be different from cluster to cluster.

With clusters of a given type, the sample-size problem is likely to be that of choosing k, the number of clusters in the group that receives a specified treatment. If $n = \Sigma n_j$ is the total number of individuals in the sample, the average size of cluster is $\bar{n} = n/k$. The most-natural estimate is usually the sample mean per individual, $\bar{\bar{y}} = \Sigma\Sigma y_{jr}/n$. Sometimes, however, it is advantageous to consider another estimate, $\bar{\bar{y}}_c = \Sigma \bar{y}_{j\cdot}/k$, the unweighted mean of the cluster means. From (4.5.1),

$$\bar{\bar{y}}_c = \mu + \frac{1}{k}\sum \gamma_j + \frac{1}{k}\sum \bar{e}_{j\cdot}$$

and

$$V(\bar{\bar{y}}_c) = \frac{\sigma_\gamma^2}{k} + \frac{1}{k^2}\sum \frac{\sigma_{ei}^2}{n_j} = \frac{1}{k}\left(\sigma_\gamma^2 + \frac{\sigma_e^2}{\bar{n}_h}\right) \tag{4.5.2}$$

if $\sigma_{ej}^2 = \sigma_e^2$ is a constant, where $\bar{n}_h$ is the harmonic mean of the n_j. A property of $\bar{\bar{y}}_c$ is that the sample variance between cluster means provides an unbiased estimate of $V(\bar{\bar{y}}_c)$, namely,

$$V(\bar{\bar{y}}_c) = \sum \frac{\left(\bar{y}_{j\cdot} - \bar{\bar{y}}_c\right)^2}{k(k-1)} = \frac{\hat{\sigma}_b^2}{k}$$

This estimate, with $(k-1)$ degrees of freedom, is unbiased regardless of whether the within-cluster variances σ_{ej}^2 vary from cluster to cluster. Thus the simple formulas of the preceding section may be used in estimating the number of clusters needed, with $\hat{\sigma}_b^2$ replacing σ^2 and k replacing the previous n. This would of course require previous data for the same type of cluster.

The result about $V(\bar{\bar{y}}_c)$ also holds when the response is a $0-1$ variate. If the estimate is the mean $\bar{\bar{p}}_c$ of the cluster mean proportions $\bar{p}_j$, the quantity $\Sigma(\bar{p}_j - \bar{\bar{p}}_c)^2/k(k-1)$ is an unbiased estimate of $V(\bar{\bar{p}}_c)$, replacing the familiar $\bar{p}\bar{q}/n$.

With the ordinary sample mean per person,

$$\bar{\bar{y}} = \frac{1}{n}\sum_j \sum_r y_{jr} = \mu + \frac{1}{n}\sum n_j\gamma_j + \frac{1}{n}\sum_j \sum_r e_{jr}$$

assuming σ_{ej}^2 constant, the variance is

$$V(\bar{\bar{y}}) = \frac{\Sigma n_j^2}{n^2}\sigma_\gamma^2 + \frac{\sigma_e^2}{n} = \left(\frac{1}{k} + \frac{\Sigma(n_j - \bar{n})^2}{n^2}\right)\sigma_\gamma^2 + \frac{\sigma_e^2}{n} \tag{4.5.3}$$

Unless the n_j vary greatly, the coefficient of σ_γ^2 is usually little larger than $1/k$.

Comparing $V(\bar{\bar{y}}_c)$ with $V(\bar{\bar{y}})$ from (4.5.2) and (4.5.3) respectively, we note that the coefficient of σ_γ^2, the between-cluster component of variance, is slightly smaller in $V(\bar{\bar{y}}_c)$, while that of σ_e^2 is slightly smaller in $V(\bar{\bar{y}})$ since $n > k\bar{n}_h$. The differences in variance are usually only moderate unless the n_j vary widely.

An unbiased sample estimate of $V(\bar{\bar{y}})$ can be constructed from an analysis of variance of the sample data. The expected values of the mean squares s_b^2 (between clusters) and s_w^2 (within clusters) work out as follows, where $Y_{j\cdot}$ is a cluster total:

$$s_b^2 = \frac{1}{k-1}\left(\sum \frac{Y_{j\cdot}^2}{n_j} - \frac{Y_{\cdot\cdot}^2}{n}\right); \qquad E(s_b^2) = \sigma_e^2 + \bar{n}'\sigma_\gamma^2$$

and

$$s_w^2 = \frac{1}{n-k}\left(\sum\sum y_{jr}^2 - \sum \frac{Y_{j\cdot}^2}{n_j}\right); \qquad E(s_w^2) = \sigma_e^2$$

where

$$\bar{n}' = \frac{1}{k-1}\left(n - \frac{\Sigma n_j^2}{n}\right)$$

(usually slightly less than $\bar{n}$). An unbiased sample estimate of $V(\bar{\bar{y}})$ is obtained by inserting s_w^2 and $(s_b^2 - s_w^2)/\bar{n}'$ as estimates of σ_e^2 and σ_γ^2 in (4.5.3). This method would enable us to attach an estimated standard error to an estimate $\bar{\bar{y}}$ from a completed survey.

In estimating k for a new survey from past data having similar clusters, one approach is to rewrite (4.5.3) in the form

$$V(\bar{\bar{y}}) = \frac{1}{k}\left[\sigma_\gamma^2\left(1 + \frac{(k-1)(CV)^2}{k}\right) + \frac{\sigma_e^2}{\bar{n}}\right]$$

where $(CV)^2$ is the square of the coefficient of variation* of the cluster sizes n_j. The quantities σ_γ^2, σ_e^2, and $(CV)^2$ could all be estimated from past data and the relation between $V(\bar{\bar{y}})$ and k estimated.

Persons unfamiliar with the implications of cluster sampling might use the estimate s^2/n for $V(\bar{\bar{y}})$, where s^2 is the usual variance between individuals in the sample. It turns out that s^2/n is an underestimate, since

$$E\left(\frac{s^2}{n}\right) = \frac{\sigma_e^2}{n} + \frac{1}{n-1}\left(\frac{k-1}{k} - \frac{\Sigma(n_j - \bar{n})^2}{n^2}\right)\sigma_\gamma^2 \qquad (4.5.4)$$

By comparison with (4.5.3) the coefficient of σ_e^2 is correct, but that of σ_γ^2 is too small, being less than $1/(n-1)$ in $E(s^2/n)$ but greater than $1/k$ in the true variance in (4.5.3). The underestimation can be serious if the clusters are large (n/k large) or if members of a cluster give similar responses, so that the σ_γ^2 dominates σ_e^2.

To illustrate, suppose $k = 10$ clusters of sizes $n_j = 10, 12, 14, 16, 18, 22, 24, 26, 28, 30$, giving $n = 200$, $\bar{n} = 20$, and $\bar{n}_h = 17.62$. We find

$$V(\bar{\bar{y}}) = 0.005\sigma_e^2 + 0.111\sigma_\gamma^2$$

$$V(\bar{\bar{y}}_c) = 0.00567\sigma_e^2 + 0.1\sigma_\gamma^2$$

*The coefficient of variation is the standard deviation divided by the mean; here it would be the standard deviation of the cluster sizes divided by their mean size.

and

$$E\left(\frac{s^2}{n}\right) = 0.005\sigma_e^2 + 0.00447\sigma_\gamma^2$$

The coefficient of σ_γ^2 in $E(s^2/n)$ is only about $\frac{1}{25}$ of the correct value in $V(\bar{\bar{y}})$. This ratio is usually near $1/\bar{n}$ (in this example $\frac{1}{20}$).

Estimation of sample size naturally becomes more difficult in samples of complex structure, since we have to develop the correct variance formula and find a previous study of the same structure with the same response variable. In agencies that regularly employ complex survey plans, a helpful device used by some writers is the following [see Kish (1965), Section 8.2]. Keep a record of the ratio of the unbiased estimate of a quantity like $V(\bar{\bar{y}})$, computed from the results of the complex survey, to the elementary but biased estimate (in this case s^2/n) given by the standard formula that assumes a random sample of individual persons. This ratio is called the design effect (deff). For designs of the same complex structure, the deff ratios are often found to be closely similar for related response variables y, y', y'', and so forth. Consequently, if a previous sample of similar structure can be found when we are planning the size of a new sample, a knowledge of these deff ratios helps in determining a realistic estimate of sample size. Even if the y variable in the new survey was not measured in the previous survey, we may be able to guess a deff ratio for this y from the ratios for related variables in the previous survey. For example, suppose that the elementary formula for $V(\bar{\bar{y}})$ suggests $n = 500$ to make $V(\bar{\bar{y}}) = 2$. If we guess a deff ratio of around 1.3 for a sample of the intended structure, we increase n to (500)(1.3) = 650.

4.6 PLANS FOR REDUCING NONRESPONSE

The term "nonresponse" is used to describe the situation in which, for one reason or another, data are not obtained from a planned member of a sample. My impression is that standards with regard to nonresponse rates are lax in observational studies; one can name major studies in which nonresponse rates of 30–40% are stated with little reported evidence of earlier attempts to reduce these high figures.

As is well known, the primary problem created by nonresponse is not the consequent reduction in sample size; this could be compensated for by planning an initial sample size larger than needed. The real problem is that nonrespondents, if they could be persuaded to respond, might give somewhat different answers from the respondents, so that the mean of the sample of respondents is biased as an estimate of the population mean. We can think of a population as divided into two classes. Class 1, with mean $\bar{Y}_1$, consists of those who would respond to the planned sample approach. The mean $\bar{y}_1$ of the sample respondents is an unbiased estimate of $\bar{Y}_1$. Class 2, with mean $\bar{Y}_2$, consists of those who would not respond. If W_1 and W_2 are the population proportions of respondents and nonrespondents, the sample estimate $\bar{y}_1$ has bias $\bar{Y}_1 - \bar{Y} = \bar{Y}_1 - W_1\bar{Y}_1 - W_2\bar{Y}_2 = W_2(\bar{Y}_1 - \bar{Y}_2)$. With this simple model the bias does not depend on the sample size, so that the bias can dominate in large samples.

Sample evidence regarding whether $\bar{Y}_1$ and $\bar{Y}_2$ are likely to differ much is of course difficult to obtain, since this involves collecting information about those who were initially nonrespondents in a sample. Such information as has been collected indicates that (1) there is usually some nonresponse bias of size $\bar{Y}_1 - \bar{Y}_2$ depending on the type of question asked and on the sample approach, and (2) the bias is not necessarily serious, but it can be. From the form of the bias, $W_2(\bar{Y}_1 - \bar{Y}_2)$, the danger of any serious bias can be kept small by keeping W_2, the nonresponse rate, small.

Fortunately, W_2 can often be materially reduced by a combination of hard work and advance planning in anticipation of a nonresponse problem. The strategy adopted for reducing W_2 will depend on one's concept of the reasons for nonresponse in a planned sample.

Consider a survey in which the approach is directly to the individual member in the sample (either by mail, telephone, or household interview). A good attitude to keep in mind is that you are asking the sample member in effect to work for you (nearly always without pay) and that the member is busy. Usually the investigator opens with a brief account of the topic of the survey, stressing its importance and the reasons why the information is needed. It is helpful to capture the respondent's interest, but this depends on the topic. Additionally, the list of questions should be designed to convince the member that you are competent and are neither wasting the member's time, prying unnecessarily, nor asking the member to respond to vaguely worded questions.

Careful thought must be given to the order in which questions are to be asked. Early questions should be important and obviously relevant to the topic. For instance, as a professor I receive questionnaires about teaching practices and about attitudes or performance of the students. If the questionnaire begins with numerous questions about my past that do not seem relevant to what the investigator has stated he is trying to learn, the probability increases that I will be a nonrespondent. The same is true if there are questions such as "How would the students react if such and such a change were made?" My reply would be "Don't ask me, ask the students," or "Don't ask anybody—the students won't know either." Answers to hypothetical questions can seldom be trusted.

Every question that the investigator proposes beyond those obviously relevant, should be justified by considering: "Is this question essential?" The investigator should know the specific role that the answer to this question will play in the analysis, and how the analysis will be weakened if this question is omitted. The nonresponse rate usually increases as the questionnaire lengthens. If a batch of questions that are needed are likely to seem irrelevant to the respondent, it may be worth inserting a brief explanation as to why these questions are essential.

Devices that save the respondent's time should be sought, for example, indicating answers to questions by placing X's in boxes instead of writing answers. Using X's is feasible only when a limited number of types of answers to a question will cover the great majority of sample members, leaving an "other" written category for those who do not choose one of the boxes. With this scheme, some pilot checking is advisable to verify that the "other" category seems small, or to add one or two additional boxes.

Give an assurance of anonymity to each sample member. Except possibly in studies in which different questionnaires are sent to the same respondents at intervals of time, the respondent's name and address may not be needed on a returned questionnaire. If not, an identifying number on a mailed questionnaire will be necessary because you will want to know the names and addresses of those who did not respond to the first mailing in order that further mailings may be made to them.

In this connection, make advance plans for a definite *call-back* or *repeated-mailings* policy on those who do not answer the first inquiry. A minimum of up to three calls is considered advisable, while high-quality studies may insist on as many as six calls, if necessary. At the same time the effect of the call-back policy on the costs and the timing of the analysis and reporting of results needs to be considered. Actually, field results show that if the cost of planning the sample and the cost of conducting the statistical analysis are included, the overall cost per completed questionnaire is little higher for a three- or even a six-call policy than for a one-call policy, but time of analysis is affected. Comparison of results for the first and each later call provides clues about the nature of nonresponse bias.

In some studies, for example, of schools or branches of a business, the situation is that if the governing bodies of these establishments are convinced of the importance of the study, they will assure that the questionnaires are answered, except for reasons such as illness. Success or failure of a study may depend largely on the amount of planning, consultation, and discussion needed in presenting the case for the study before these governing bodies. The nonresponse problem here occurs in lumps; that is, refusal by a governing body may mean that 10 or 20% of the sample is missing in one decision. Persuading such a governing body to change its mind (the analogy of a successful call-back) challenges the ingenuity of the investigator.

4.7 RELATIONSHIP BETWEEN SAMPLED AND TARGET POPULATIONS

At some point in the planning it is well to summarize one's thinking about the relationship between the sampled populations—the populations from which our comparison groups will be drawn—and the target population to which we hope that the inferential conclusions will apply. Availability and convenience play a role, sometimes a determining role, in the selection of sampled populations. On reflection these may be found to differ in some respects from the target population. This issue is not confined to observational studies. For instance, controlled experiments in psychology may be confined to the graduate students in some department or to volunteer students at $5 per hour, although the investigator's aim is to learn something about the behavior of graduate students generally or even of all young people in this age range in the country. Airline pilots might be a convenient source of data for an inexpensive study of men's illnesses in the age range 40–50, but we hope that they are not typical of men, generally, in the frequency or severity of strokes or heart attacks.

The problem is that results found in the sampled populations may differ more or less from those that would be found in the target population. In studies of the economics of farming, a good source is a panel of farmers who regularly keep careful records of their economic transactions, in cooperation with a state university. But as would be expected, there is evidence (Hopkins, 1942) that such farmers receive a higher economic return from capital invested on their farms and adopt improved techniques more rapidly than do farmers generally.

This issue is common in program evaluation also, for example, teaching programs or client-service programs. Owing to difficulty of taking accurate research measurements within an operating program, it may be decided to conduct the study *outside* the program, although its results are intended to apply to the program. The change in the setting can affect the results. If the study is conducted *inside* the program, workers in the program, aware that they are being tested, may perform better in the study than they usually do in the ordinary operation of the program. Alternatively, the study, if imposed as a temporary extra load of work, might result in a lower quality of performance than is regularly attained. If a change in procedure in the program has been decided, the old and the new procedure may both be continued for a time in order to measure the size of the presumed benefit from the change. In this event, the workers, aware that the old procedure is to be abandoned, may do only slipshod work on it during the trial period, resulting in an overestimate of the benefit, if any, from the change.

The particular years in which the study is done may affect the results. A comparison of public versus slum housing might give one set of results in a period of full employment and rising prosperity, during which the slum families have the resources to move to superior private housing, but a different set if unemployment is steady and money is scarce. A well-planned series of experiments on the responses of sugar beet to fertilizers at the major centers in England was conducted at about 12 stations each year. After three years an argument arose for stopping the experiments because effects, while profitable, had been rather modest from year to year; the average responses to 90 lbs. nitrogen per acre were 78, 336, and 302 lbs. sugar per acre in the three years. There seemed little more to be learned. A decision to continue the experiments was made, however, because all three years had unusually dry summers. In the next two years, both wet years, the average effects of nitrogen rose to 862 and 582 lbs. sugar per acre.

Reflection on likely differences between sampled and target populations has occasionally caused investigators to abandon a proposed plan for a study, because the only available locale seemed so atypical of the target population that they doubted whether any conclusions would apply. Sometimes, the choice between two locales, otherwise about equally suitable, was made on this criterion. If resources permit, it might be decided to conduct the study in each of two locales that were atypical in different respects, or to have two or three control groups instead of only one, for instance where a new urban-renewal program is being tried in one town, and the "control" has to come from neighboring towns.

Supplementary analyses may help in speculating whether results obtained in one population are likely to hold up in another population. For instance, since different populations usually show somewhat different distributions of ages, economic levels, sex ratio, and urban–rural ratio, a statistical examination is relevant for this problem in a study that reveals the extent to which an estimated treatment effect varies with the level of any of these variables. The investigator might well regard it as part of his/her responsibility to report any aspect of the results or any feature of the sampled population that is similarly relevant. Research data on methods of handling some important social problems are scarce. An administrator in Washington, D.C. or in California is likely to use the results of any study that can be found for policy guidance, for example, a study conducted on a particular group of people in Manhattan.

4.8 PILOT STUDIES AND PRETESTS

Early in the planning the investigator should begin to consider what can be learned from pilot studies and pretests. Most written discussions of the role of pilot studies deal with household-interview surveys (and to some extent with telephone or mail surveys), in which information can be gained about such issues as:

1. Ability of the interviewers to find the houses.
2. Adequacy of the questionnaire; for example, do some questions elicit many refusals, or many "don't knows" that could perhaps be avoided by a change in the form or the ordering of the questions.
3. Some indication of nonresponse rates.
4. A check on advance estimates of time per completed questionnaire per house and of costs of the field work.
5. Trial training for the interviewers.
6. If the planning team is undecided in selecting alternative forms of some questions, the effects of different question ordering, or household interviews versus telephone interviews with household interview used only in follow-up, a proposed pilot sample can be divided into random halves, using one alternative in each half.

One question is: Need the pilot sample be a random subsample of the whole planned sample, versus one chosen for speed and convenience in an area easily accessible to the planning headquarters? My opinion leans toward the latter choice, provided that the chosen pilot sample is judged to be reasonably representative of the range of field problems; for instance, we would obviously not want a pilot sample confined to rich person's houses if the questionnaire problems are likely to be relevant among the poor. If the pilot sample were intended to estimate variability for determination of sample size, it would have to be a random subsample, but pilot samples are used for this purpose only in planning major and expensive surveys in which substantial time and resources for a pilot study are considered essential. Parten (1950) and Moser (1959) provide good references to the roles of pilot samples in surveys.

If an observational study is to be conducted from existing records on individual persons, originally collected for another purpose, a pilot study of these records is highly advisable before committing oneself to the proposed study. Items to check include completeness, signs of gross errors, understanding of definitions, and signs of subtle changes in the meaning of the terms over time—or more generally, to provide an appraisal of the quality of the records for the intended purpose. It may be quite convenient to draw, say, an "every *k*th" systematic sample with a random start for the pilot. Time must be provided for the necessary statistical analysis of this pilot and for attempted follow-up of signs that arouse suspicion.

In a study to be done from state, regional, or national summary data, one goal of pilot work is to learn as much as possible about completeness of the data and any known or suspected biases. Useful strategies include discussions with persons involved in the collection of these data, searches for critiques of the data by outside persons, and preliminary graphical analyses (e.g., looking for sudden departures from smooth curves) followed by further discussion.

Since observational studies vary widely in nature, a further listing of possibilities will not be attempted. Try to plan any pilot work to aid a specific decision about the conduct of the study, rather than just having a look at how things go.

4.9 THE DEVIL'S ADVOCATE

When the plans for the study near completion, consider presenting a colleague with a fairly detailed account of the objectives of the study and your plans for it, and ask that person to play the role of devil's advocate by finding the major methodological weaknesses of your plan. It may be difficult to persuade this person to do this, since you are usually requesting more than a trifling amount of work. On the other hand, some scientists enjoy criticizing another's work and are good at it.

I stress this point because most observational studies, particularly those of any complexity, have methodological weaknesses. Some weaknesses are unavoidable due to the nature of an observational study or to the types of comparison groups available to us. Some weaknesses could be removed either by collecting data that we did not intend to collect at first or by using a more searching statistical analysis, as is seen when the investigator tries to reply to a slashing critique of his/her results that appears after the study has been published. Some readers may see the critique, but may see neither the investigator's original study nor the rebuttal. For some weaknesses that cannot be removed from the study, additional data or analyses may enable the investigator to reach a judgment regarding the strength of the objection, which the investigator can then publish as part of the report. Published results of studies on current social problems often rouse emotional reactions for or against the conclusions of the investigator. The reader who is emotionally against the investigator's conclusions is apt to magnify any criticism of a study that appears later. For this reason, a good practice in reporting is to list and discuss any methodological weakness or possible objection that has occurred to the investigator.

4.10 SUMMARY

Statistical theory provides formulas that aid in the estimation of the size of samples needed in a study. These formulas are worth using even when the choice of sample size is dominated by considerations of cost or number of available subjects. For exploratory studies, where the issue is whether a given agent or treatment produces any effect, the formulas estimate the sample size that will ensure a high probability of finding a statistically significant effect of the treatment when the real effect has a specified size δ. If the objective is estimation, the formulas give the size needed to make the estimate correct to $\pm L$ with high probability. Illustrations of the uses of the formulas in some simple problems are presented.

In observational studies the primary difficulties in using the formulas are (1) estimates of population parameters that appear in the formulas must be inserted, (2) biases increase the type-I errors in tests and decrease the probability that the estimated treatment effect is correct to within the stated limits, (3) the samples are often of more-complex structure than assumed in the simple formulas. An example of this type is given in which the sample consists of groups or clusters of subjects, rather than individual subjects.

The term "nonresponse" is given to the failure to obtain some of the planned measurements. The problem with nonresponse is not so much the reduction in sample size, which can be compensated for by a planned sample larger than is needed. There is, however, evidence that people unavailable or unwilling to respond may differ systematically from those who respond readily, so that results from the respondents are biased if applied to the whole population. In planning, likely sources of nonresponse need to be anticipated and plans need to be made to keep the level of nonresponse low. Repeated call-backs are a standard device in mail and household-interview surveys. Questionnaires should be constructed so as to gain the respondent's interest, respect, and confidence. Sometimes, the main hurdle is to devise an approach that will obtain permission and support from an administrative or governing body.

In both observational studies and controlled experiments, the population represented by the study samples may differ from the target population to which the investigator would like the results to apply. At some point in the planning, the investigator should reflect on the differences between the sampled and target populations; sometimes, supplementary analyses can be carried out that help in judging to what extent results for the sampled population will apply to the target population.

The reasons for a pilot study on some aspect of the proposed plan should be considered. In an interview survey, pilot studies can gain information on such matters as wording, understanding and acceptability of the questions, the sources and nature of the nonresponse problem, the time taken, and field costs. In a study of existing records, completeness and usability of the records for research purposes can be checked, and, more generally, any uncertain aspect of the proposed plan.

When the proposed plan nears completion, a colleague capable of critiquing the plan can help by reviewing the plan, pointing out methodological weaknesses that have escaped the notice of the planners, and, if possible, suggesting means of remedying these weaknesses. The report of the results should discuss weaknesses that cannot be removed from the plan, and give the investigator's judgment regarding the effects of the weaknesses on the results.

REFERENCES

Hopkins, J. A. (1942). Statistical comparisons of record-keeping farms and a random sample of Iowa farms for 1939." *Agr. Exp. Sta. Res. Bull.*, **308**, Iowa State College.

Kish, L. (1965). *Survey Sampling.* Wiley, New York.

Moser, C. A. (1959). *Survey Methods in Social Investigation.* Heinemann, London.

Parten, M. B. (1950). *Surveys, Polls and Samples: Practical Procedures.* Harper & Brothers, New York.