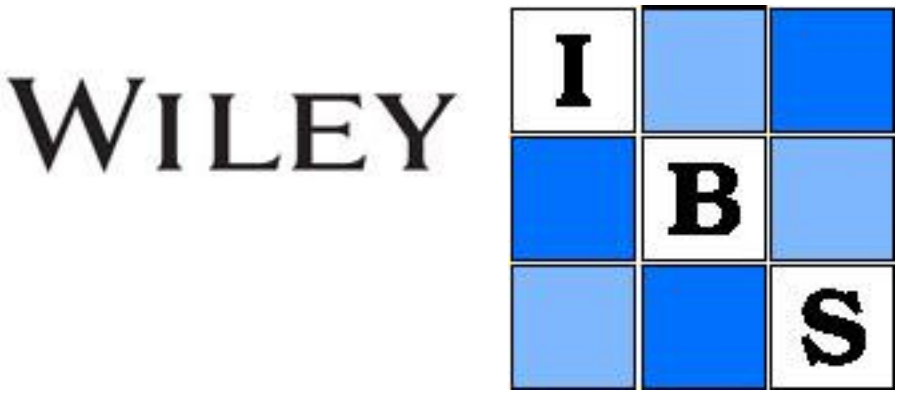

FIXED-SAMPLE-SIZE ANALYSIS OF SEQUENTIAL OBSERVATIONS

F. J. ANSCOMBE
Statistical Laboratory, Cambridge, England

The methods most commonly employed for the statistical analysis of observations are based on the assumption that the number of observations was decided on in advance. The number of observations is indeed chosen in advance in many types of experiment or observational inquiry. In an agricultural field experiment, the number of plots and their treatments must be completely specified long before any observations can be taken; and (apart from possible failures which will be recorded as "missing observations") the number of observations eventually obtained is precisely the number chosen at the outset. Many chemical determinations are made in triplicate or quadruplicate or some other fixed number of times, according to a definite rule; and so the number of readings obtained in each determination is fixed in advance. A sample survey of households in a town may be based on inquiries at every hundredth house in a list; here again the number of observations is fixed in advance, provided we include the instances of non-response.

There are other sorts of inquiry where the number of observations is not fixed in advance, and where the experimenter, if asked just before he began, would not be able to say how many observations he would take. He may, for example, be following some recognized type of sequential sampling rule, such as one of A. Wald's sequential tests, or the inverse sampling of J. B. S. Haldane and M. C. K. Tweedie. In that case, when reporting his observations he will presumably state what the sampling rule was, and he will use a method of statistical analysis specially designed for that sampling rule.

Most commonly, however, when the number of observations is not fixed in advance, this is because at the outset the experimenter has not fully made up his mind as to his requirements or resources or the nature of the material being studied; and so he does not decide to make a certain fixed number of observations nor to follow a definite sequential sampling rule, but proposes simply to take observations until such time as it shall seem appropriate to stop. Sometimes indeed

the experimenter does follow a definable sampling rule, but as it is not of an orthodox sequential type, he does not think of it as "sequential". It is convenient to call any such observations, taken successively and with the total number not preassigned, *sequential observations*, even though the sample size was not determined by a well-defined sequential sampling rule consciously adopted by the experimenter. Such observations are sometimes reported without any statement as to how their number was determined, the experimenter considering this to be an irrelevant detail; and it is usual to treat the observations in the statistical analysis as if their number had been fixed in advance. The purpose of this paper is to consider a number of different situations of this sort, to see whether any serious error will result from a fixed-sample-size analysis. The paper is a continuation and partial summary of two recent discussions in [2] and [4], where bibliographies of relevant literature may be found.

Sample Size Independent of the Observations

Sometimes it happens that, although the sample size is not fixed in advance, and depends on various circumstances, it does not depend at all on the observations themselves. For example, the observer may decide to take as many observations as he can in a limited period of time. Provided the time taken over an observation is not correlated in any way with the reading obtained, the decision to stop does not depend on the previous readings. In an investigation of fleas on rodents, the observer may decide to catch as many rodents of each species as he can, but to discontinue catching any species of which he has already caught 100 specimens. With such a sampling procedure, the number of animals of any one species caught is an uncertain quantity, but (presumably) the number does not depend on the flea populations of the animals caught.* The situation would be quite different if the sampling

*I am suggesting that probably in such a situation the number of animals caught would not be correlated with the flea populations of the animals caught. This would be the case if the activity of the animals was not in any way affected by the number of fleas carried. It might still be the case even if the number of fleas affected activity; for example, if the more heavily infested animals were less active and less likely to be caught. Then the sample of animals obtained by trapping would be biased in favor of less heavily infested animals, and this would be so whether the number of animals to be caught was fixed in advance or determined by the sampling rule stated. The bias would be a property of the trapping method, not of the statistical sampling rule; and there might in fact be no correlation of the type under consideration, even though there was a bias. But (just as in the following example of a sample survey of households) one can imagine circumstances that could produce such a correlation. For example, if heavily infested animals were gregarious and likely to be caught in a bunch, if at all, while the other animals were not gregarious, and if heavily and lightly infested animals did not interfere with each other, then under the sampling rule stated there would be a positive correlation between the total number of animals caught and the infestations of the animals caught. I presume that an effect of this sort is in fact unlikely.

procedure were modified so that the observer left off catching a species as soon as he had caught 100 *infested* specimens; for now the number of animals caught would depend on how many of the caught animals were not infested.

As another example, consider again the sample survey of households in a town, where the total number of houses visited was fixed in advance. The number of houses from which satisfactory replies are obtained is an uncertain quantity, but in ordinary circumstances this number will not be correlated in any way with the replies themselves. Thus if we ignore the non-responses, we can say that the number of observations (i.e. of responses) is independent of the observations. One can, however, imagine a situation in which that is not the case, and I think it is instructive to consider such a situation, to see that the circumstances required are rather unusual. Suppose that one of the items of information being sought in the survey is the occupation of the householder, and suppose that certain questions asked in the inquiry would be deemed mischievous by persons of a certain occupation, let us say by the bookmakers, if the inquiry were brought to their attention. Then the inquiry would proceed normally unless and until the house of a bookmaker was visited. In that event, the householder, while giving the required information, would become incensed, and would forthwith start a campaign in the press and elsewhere to urge people not to cooperate in the inquiry, on some suitable grounds. It might be found that for the remainder of the inquiry the proportion of non-responses was much higher than it had been before. Then the number of satisfactory responses would depend on the responses themselves, since it would depend on whether the bookmaking profession was included among the occupations observed.

In any experiment or sampling inquiry where the number of observations is an uncertain quantity but does not depend on the observations themselves, it is always legitimate to treat the observations in the statistical analysis as if their number had been fixed in advance. We are then in fact using perfectly correct conditional probability distributions. The possibility of error in using a fixed-sample-size analysis of sequential observations therefore only occurs when the number of observations depends on the observations themselves. There are various ways in which such a dependence can be produced. I have already mentioned two examples (both akin to ordinary inverse binomial sampling; they can be described roughly as sampling to obtain not more than 100 infested animals, and not more than one bookmaker, respectively). I now consider some further examples in greater detail.

Sampling to Reach a Foregone Conclusion

Suppose that someone wishes to prove that a coin is biased. He may adopt the following procedure. He spins the coin repeatedly and keeps count of the numbers of heads and tails. He stops sampling when first the difference between the cumulated numbers of heads and tails is significant at some preassigned significance level, let us say at the 5% level, when the difference is tested as for a fixed sample size by reference to a binomial distribution with equal probabilities for heads and tails. It can be proved (with the aid of Khintchine's iterated logarithm theorem) that sooner or later a "significant" difference will be observed, so that the rule can in fact be followed; there is no question of having to go on forever. After any number n of spins, let x be the cumulated number of heads and y the cumulated number of tails ($x + y = n$). The progress of the sampling can be represented on a diagram by the locus of a "sample point" with Cartesian coordinates (x, y). Sampling terminates as soon as an appropriate boundary point is reached. The first few boundary points are shown as heavy dots in Fig. 1; for large n the boundary points lie approximately on the parabola

$$x - y = \pm 1.9600\sqrt{x + y}.$$

Clearly the use of the ordinary fixed-sample-size significance test at the end of this sampling procedure is completely invalid; for the probability of obtaining a significant result, if the null hypothesis that the coin is unbiased is true, is not just under 5% but is actually 100%.

One can similarly frame a sampling procedure designed to support the hypothesis that the coin is unbiased. It is necessary now to stop sampling *before* the cumulated numbers of heads and tails differ significantly at the chosen level when tested as if the sample size were fixed. One such procedure is represented by the boundary points shown as small circles in Fig. 1. For sample sizes less than 25, the boundary is chosen so that the difference between the cumulated numbers of heads and tails is not significant at the 10% level according to the usual test (without randomization device), but so that the difference *could* be significant at the 10% level if one further observation were taken. Thus as large a sample size as possible, but with upper limit at 25, is taken, consistent with certainly not obtaining any difference significant at the 10% level.

Here again the use of the ordinary fixed-sample-size test is completely invalid, for the probability of obtaining a result significant at the 10% level, if the null hypothesis that the coin is unbiased is true, is not just under 10% but is actually zero.

I have spoken of spinning a coin to see whether it was unbiased, but I could equally well have referred to testing the sex-ratio in an animal population, or to testing anything else that can be regarded as a binomial probability. The sampling procedure of continuing to take

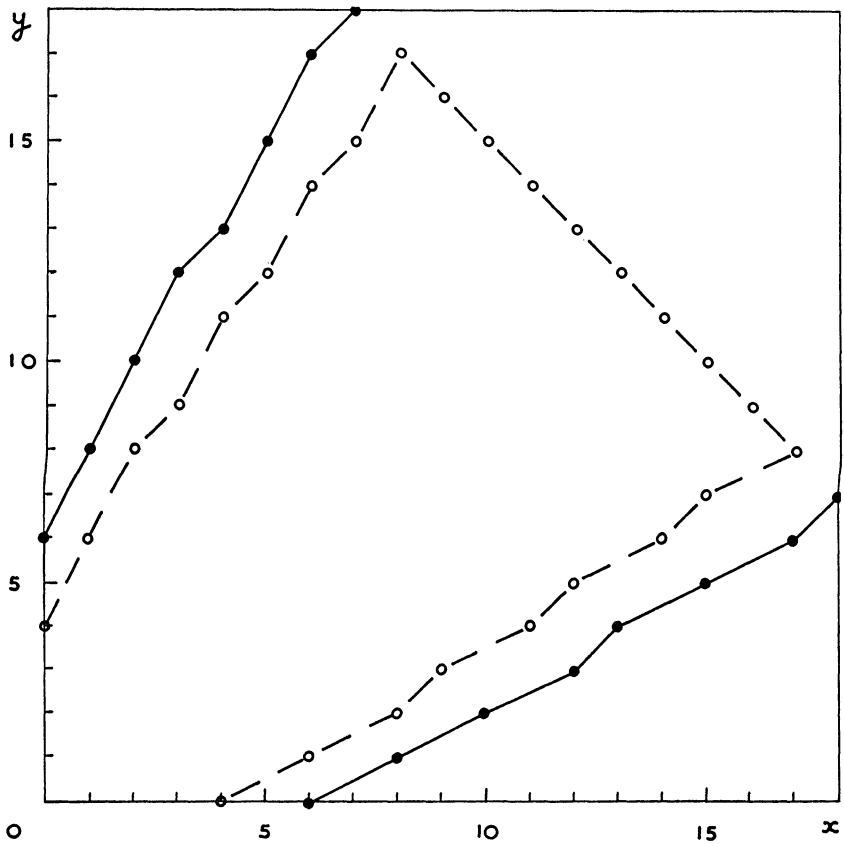

FIGURE 1. SPINNING A COIN TO REACH A FOREGONE CONCLUSION

Abscissa: cumulated number of heads.
Ordinate: cumulated number of tails.

The heavy dots are the first part of a boundary designed to demonstrate that the coin is biased. The small circles constitute a boundary designed to demonstrate that the coin is not biased.

observations until a result significant at a preassigned significance level has been obtained can, I think, be followed with any ordinary significance test whatever. Robbins [5] has discussed a procedure for showing that the mean of a normal population differs significantly from a stated

value; it is exactly parallel to that just considered for a binomial population.*

If an experimenter deliberately follows one of these procedures for reaching a foregone conclusion, and does not admit it when reporting his work, he can reasonably be accused of dishonesty. But it can happen that something of the sort occurs without any conscious intention to deceive. There has been a good deal of discussion of such questions in the literature on extra-sensory perception. In particular, Feller [3] has given a careful and interesting account of the possible effects of "optional stopping", and also of the non-publication of results that are considered to be uninteresting.

Double Sampling

It sometimes happens that an experimenter begins by taking a certain number of observations, and then, if he considers that the results are interesting he takes some more observations (perhaps as many again), but if the results are not interesting he does not take any more observations. In the comparison of a new treatment with an old one, if the first sample indicates fairly clearly that the new treatment is not superior to the old one the experimenter will probably wish to discontinue the test, but if it seems possible that the new treatment is superior to the old he will wish to investigate the matter further. At the end, he will no doubt pool all the results and treat them as if the sample size had been fixed in advance.

To see how great an error may be committed in such an analysis, let us consider the following specific problem. The mean μ of a normal population is to be estimated. The variance σ^2 of the population is supposed known (or the sample is large enough for it to be estimated with negligible error). A first sample of n_1 observations is taken and its mean $\bar{x}_1$ is calculated. If $\bar{x}_1 > 0$, a second sample of n_2 observations is taken, while if $\bar{x}_1 \leq 0$, no further observations are taken. (Here n_1 and n_2 are fixed.) Let N denote the total number of observations (so that $N = n_1$ or $n_1 + n_2$) and $\bar{x}$ the average of all the observations (we have $\bar{x} = \bar{x}_1$ if $\bar{x}_1 \leq 0$). An attempt is made to give 95% confidence limits for μ by calculating

$$\bar{x} \pm 1.9600\sigma/\sqrt{N}.$$

What is the true probability that μ lies between these limits, for any fixed value of μ?

*Robbins was interested in cure as well as diagnosis. Having pointed out the possibility of sampling to reach a foregone conclusion, he showed how the nominal significance level of the fixed-sample-size test might be adjusted to allow for optional stopping anywhere between two preassigned sample sizes.

It turns out that the answer depends on the values of $(\mu\sqrt{n_1})/\sigma$ and of n_2/n_1 . The biggest variations occur when n_2 is much larger than n_1 . Fig. 2 shows how the true probability varies with μ in the mathematically-simple limiting case when n_2/n_1 is infinite (continuous curve), and also in the case $n_2 = n_1$ (broken curve). The ordinate is

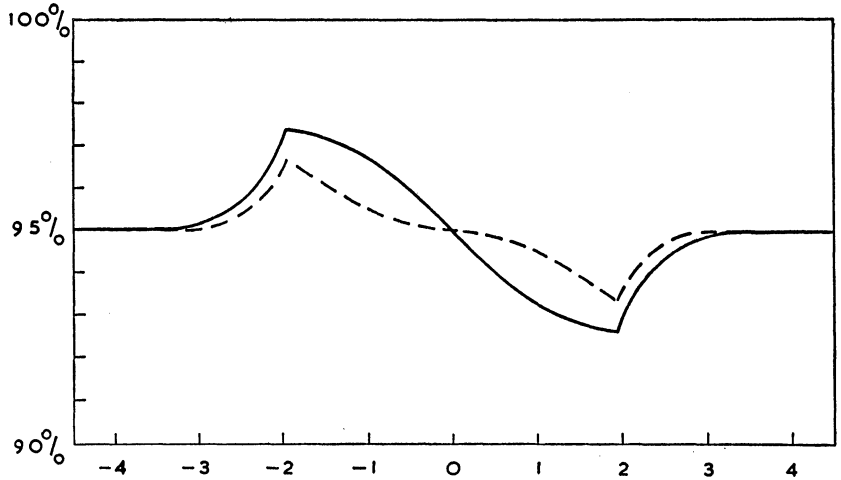

FIGURE 2. DOUBLE SAMPLING TO ESTIMATE THE MEAN OF A NORMAL POPULATION

Abscissa: value of $(\mu\sqrt{n_1})/\sigma$.
Ordinate: probability that μ lies between the calculated limits.
The continuous curve is for n_2/n_1 infinite. The broken curve is for $n_2 = n_1$

the probability that μ lies between the calculated limits, the abscissa is the value of $(\mu\sqrt{n_1})/\sigma$. For n_2/n_1 infinite, the probability varies between 92.625% (realized when μ is such that the probability that only one sample will be taken is just 2.5%) and 97.375% (realized when μ is such that the probability that only one sample will be taken is just 97.5%). If μ is so large that it is almost certain that two samples will be taken, or so small that it is almost certain that only one sample will be taken, the probability that the calculated limits will bracket μ is very close to the intended value of 95%.

For smaller ratios of n_2 to n_1 , the deviations of the true probability from 95% are reduced in magnitude. Limits between which the true probability varies are approximately as follows:—

For $n_2 = \frac{1}{2}n_1$: 93.65% and 96.35%.

For $n_2 = n_1$: 93.34% and 96.66%.

For $n_2 = 2n_1$: 93.08% and 96.92%.

The values of $(\mu\sqrt{n_1})/\sigma$ at which these extreme values are attained are the same as before, namely ± 1.9600.

I have supposed that a second sample would be taken whenever the mean of the first sample exceeded a certain value (taken to be 0). Sometimes a second sample is taken only if the mean of the first sample is neither very low nor very high, say only if $\bar{x}_1$ lies between $-a$ and a (a being a fixed positive quantity). The disturbances produced in the true probability that μ lies between the calculated confidence limits, due to the abrupt changes in N for $\bar{x}_1$ near to $-a$ and near to a, are independent and additive; and the greatest divergence between the true and nominal probabilities occurs for $(a\sqrt{n_1})/\sigma = 1.9600$, n_2/n_1 large, and $\mu = 0$, when the true probability is 90.25%.

We can rephrase this last result as follows. A first sample of fixed size n_1 is taken. If the mean $\bar{x}_1$ does not differ significantly at the 5% level from some critical value, say 0, a second sample of fixed size n_2 is taken, n_2 being much larger than n_1 . Otherwise, no further observations are taken. Then if the population mean is in fact 0, the probability that the mean of all the observations will differ significantly from 0 at the 5% level, according to the ordinary fixed-sample-size test, is not 5% but 9.75%.

A Confidence Interval of Preassigned Width

A refinement of the preceding double-sampling procedures is to take observations in several small samples, or even one by one, until the required amount of information has been obtained. To see the effect of this greater flexibility of sampling, let us consider the following specific example. It is required to estimate the mean μ of a normal population by a confidence interval of width l and coefficient $1 - \alpha$, the population variance σ^2 being unknown. When any number n of observations have been taken, confidence limits for μ with coefficient $1 - \alpha$ are given, according to the usual fixed-sample-size theory, by

$$\bar{x}_n \pm s_n t_\alpha^{(n-1)}/\sqrt{n},$$

where $\bar{x}_n$ is the mean of the n observations, s_n^2 is the usual unbiased quadratic estimate of σ^2, and $t_\alpha^{(n-1)}$ is such that a random variable following Student's distribution with $n - 1$ degrees of freedom has probability $1 - \alpha$ of lying between $\pm t_\alpha^{(n-1)}$. Consider the sequential sampling rule: observations are taken one by one until first

$$s_n t_\alpha^{(n-1)}/\sqrt{n} \leq \tfrac{1}{2}l. \tag{1}$$

Denoting this value of n by N, we ask what is the true probability that μ lies between $\bar{x}_N \pm \frac{1}{2}l$.

Provided l is small, so that N is large, it can be shown that this confidence interval has very nearly the required confidence coefficient of $1 - \alpha$. Let

$$\nu = \frac{4\sigma^2 t_\alpha^2}{l^2},$$

where t_α is such that a standard normal variable has probability $1 - \alpha$ of lying between $\pm t_\alpha$. When it is an integer, ν is the number of observations that would be taken if σ^2 were known beforehand. Let $\varphi(t_\alpha)$ denote the ordinate of the frequency function of a standard normal variable when the abscissa is t_α. Then it can be shown (by an easy deduction from formulae given in [2]) that the true probability that μ lies between $\bar{x}_N \pm \frac{1}{2}l$ is approximately

$$1 - \alpha - \frac{1.176 t_\alpha \varphi(t_\alpha)}{\nu}, \tag{2}$$

provided ν is large (and provided that N is in no case allowed to be less than 4). If for example $\alpha = 5\%$, the true confidence coefficient is approximately

$$95 - \frac{13.5}{\nu}\,\%. \tag{3}$$

If instead of using the percentage point $t_\alpha^{(n-1)}$ of Student's distribution in the sampling rule (1) we had used the percentage point t_α of the standard normal distribution, (3) would have been changed to

$$95 - \frac{41.2}{\nu}\,\%. \tag{4}$$

Thus the error resulting from treating the sequential observations as if the sample sized were fixed is of the same order of size as, but less in magnitude than, the error in replacing Student's distribution by a standard normal distribution. Clearly, for practical purposes, the error is negligible unless the average sample size is quite small.

A similar procedure can be followed when estimating the difference between the means of two normal populations of which the variances are unknown but supposed to be equal. The observations are taken in pairs, one from each population, and after n pairs have been taken an estimate of the population variance is available having $2(n - 1)$ degrees of freedom. Sampling continues until the confidence interval with coefficient $1 - \alpha$ for the difference in means, calculated according to the usual fixed-sample-size formula, is first less than or equal to the given length l in width. If $\bar{y}_N$ denotes the difference in sample means

when sampling terminates, the confidence interval is taken to be $(\bar{y}_N - \frac{1}{2}l, \bar{y}_N + \frac{1}{2}l)$. Let

$$\nu = \frac{8\sigma^2 t_\alpha^2}{l^2},$$

where σ^2 denotes the common population variance. Then the true confidence coefficient for the calculated interval is approximately

$$1 - \alpha - \frac{0.255 t_\alpha \varphi(t_\alpha)}{\nu}, \tag{5}$$

provided ν is large (and provided that N is in no case allowed to be less than 3). For $\alpha = 5\%$, this is

$$95 - \frac{2.9}{\nu}\,\%. \tag{6}$$

The divergence from the intended figure of 95% is considerably smaller than that indicated before at (3). If the percentage point of Student's distribution is replaced by the standard normal percentage point in the stopping rule, the result corresponding to (4) above is

$$95 - \frac{16.8}{\nu}\,\%. \tag{7}$$

It is not the purpose of this paper to discuss the specification of sampling rules which will lead exactly or almost exactly to estimates having preassigned properties; I am concerned merely to show what error results from the strictly incorrect use of a fixed-sample-size formula. But it is perhaps worth while to point out that the sampling rule considered above for estimating the mean of a single normal population can be improved very easily as follows. Observations are taken one by one until first the inequality (1) is satisfied, and then one further observation is taken. Denoting the final number of observations by N, we calculate $\bar{x}_N \pm \frac{1}{2}l$. The probability that μ lies between these limits is now given approximately by the expression (2), except that the factor 1.176 is replaced by 0.176. The error is thus about one seventh of its previous value, and in (3) the factor 13.5 becomes 2.0. No device of this sort leads to any improvement in the sampling rule given above for estimating the difference in means of two normal populations, but the error in that case is already very small. (I am indebted to Professor J. W. Tukey for the suggestion that a simple improving device should be sought.)

Discussion

We have seen that, when the number of observations depends on the observations themselves, it can happen that a fixed-sample-size analysis of the observations is grossly wrong (as with our examples of sampling to reach a foregone conclusion), and it can also happen that a fixed-sample-size analysis is only very slightly in error (as with our examples of sampling to obtain a confidence interval of preassigned width). The examples of double sampling considered were intermediate, in that for some values of the unknown parameter there might be an appreciable (but not gross) error in the fixed-sample-size analysis, while for other values the error was negligible.

The magnitude of the possible error in using a fixed-sample-size method of analysis is related to the dispersion of the sample size. By "dispersion of the sample size" I mean the variability in sample size that would be observed if the experiment were repeated several times under similar conditions. It has been shown that if the average (or median) sample size is large and if the dispersion of the sample size is relatively small (say, if the coefficient of variation of the sample size is small), then there will be little error in treating the observations as if the sample size were fixed (see [1]). These conditions are fulfilled in the examples of sampling to obtain a confidence interval of preassigned width l, provided l is small. If the sampling were carried out several times, the sample sizes would probably show only a very small percentage variation. The conditions are not fulfilled in the examples of double sampling, unless the probability is either close to 0 or close to 1 that only one sample will be required. When $n_2 = n_1$, the coefficient of variation of the sample size can be as high as 35%, this being the value when the probability that only one sample will be taken is about 2/3; if $n_2 = 2n_1$, the coefficient of variation can be as high as 57%; and so on. These are not very small coefficients of variation. In the examples of sampling to reach a foregone conclusion the sample size has a very great dispersion, both very low and very high values being quite probable. If the procedure represented by heavy dots in Fig. 1 is followed, using an unbiased coin, the distribution of sample size has an infinite mean, and yet there is a substantial probability (about 0.11) that the sample size will not exceed 20.

Thus we may suspect appreciable error in a fixed-sample-size analysis if the following conditions are both satisfied:—

(i) *the number of observations depends on the observations themselves,*
(ii) *the relative dispersion of the number of observations in repeated sampling is not very small.*

But even if these conditions are satisfied, there is not necessarily any danger of error. It is only with certain sorts of statistical analysis that we can be misled in treating the sample size as fixed, namely when

(iii) *reference is made in the statistical analysis to some property of the distribution of a statistic.*

Methods of analysis of this sort include: (1) calculating an unbiased estimate, with or without its standard error, (2) calculating a confidence interval, in the sense of J. Neyman's theory, (3) making a significance test, in the sense of the Neyman-Pearson theory of tests. The terms "unbiased", "standard error", "confidence coefficient", "significance level", "critical region", etc., can be explained only by reference to the sampling distribution of a function of the observations; and it is because this sampling distribution (given the sample size) is liable to be affected by the sequential sampling rule followed that we run a risk of error in supposing that the sample size is fixed when it is not.

All risk of error is avoided if the method of analysis uses the observations only in the form of their likelihood function, since the likelihood function (given the observations) is independent of the sampling rule. One such method of analysis is provided by the classical theory of rational belief, in which a distribution of posterior probability is deduced, by Bayes' theorem, from the likelihood function of the observations and a distribution of prior probability. Closely related to this is R. A. Fisher's method of fiducial inference. Pragmatic methods of analysis in which the expected risks attached to alternative decisions are considered are also based on likelihoods, namely Wald's method of minimax decision functions and various developments and modifications of that, in particular D. V. Lindley's method of minimum unlikelihood.

Unfortunately many of us find the sort of analysis that refers to the distribution of a statistic more serviceable in practice than methods based directly on likelihood, despite the theoretical advantages of the latter. The topics of this paper therefore seem to me to be worth occasional consideration. If we are able to conclude that the dangers feared are unimportant and negligible in the situations with which we have to deal, then so much the better.

REFERENCES

[1] Anscombe, F. J. Large-sample theory of sequential estimation. *Proc. Camb. phil. Soc.*, 48 (1952), 600.
[2] Anscombe, F. J. Sequential estimation. *J. R. statist. Soc.* B, 15 (1953), 1.
[3] Feller, W. K. Statistical aspects of ESP. *J. Parapsychol.*, 4 (1940), 271.
[4] Lindley, D. V. Statistical inference. *J. R. statist. Soc.* B, 15 (1953), 30.
[5] Robbins, H. Some aspects of the sequential design of experiments. *Bull. Amer. math. Soc.*, 58 (1952), 527.