FORECASTING AND ESTIMATION OF CROP YIELDS

By V. G. PANSE[1] and R. J. KALAMKAR[2]

FORECASTS and estimates of yield of commercial crops like cotton, jute or sugarcane are of considerable importance to the trade and industry, because the availability of raw materials during the season is the basis of all calculations of manufacturing processes. With the increasing emphasis on 'planned' production, a still greater value will come to be attached to reliable estimates of yield, while in an emergency, like the present, arising out of war conditions, accurate forecasts and estimates of production are a paramount need for ensuring sufficiency of food grains and their equitable distribution in different areas. In India, where tax on agricultural land forms the principal source of Government revenue, the Government administration is specially interested in forecasting and estimation of crop yields.

Forecasts or estimates of most probable production are made while the crop is still standing in the field, whereas the actual production is estimated at or soon after harvest. The latter may be treated as a more accurate forecast concerning the movement and arrival of the crop to the market during the season. Estimation of production involves a knowledge both of the average yield per acre and of the total acreage sown with the crop. In England and the United States of America crop forecasts are made by a large number of voluntary reporters who are in close touch with the farming of their respective areas. In America, crop reporters are required to estimate both the yield per acre and the acreage under the crop; but in England acreage figures are obtainable precisely, since compulsory returns for all holdings are made to crop reporters. Except in permanently settled districts, an elaborate Government organization extending to the remotest village looks after crop forecasting in India. Each village has a *patwari* or an accountant who is also the crop reporter. His estimate of seasonal yield is usually expressed as a fraction of the normal yield; a method similar to that adopted in America. Acreage figures are recorded in the village register, which contains a list of all fields in the village, their dimensions and areas sown in each field with different crops each season. Area figures for different crops are consequently known with a high degree of accuracy.

The chief defect of the present methods of forecasting yield in India as well as in other countries is that no objective procedure is employed in arriving at the estimates which are merely opinions of individuals as to what the yield is. The normal yield, which forms the basis, has no precise definition. In America it is understood to represent yield better than the average but less than the maximum. In India, a certain number of crop cutting experiments are conducted on selected plots of land; but a straight average of these experiments is not taken as the normal yield. The figure adopted is based on selected results coupled with local information and opinions of revenue officials. In the absence of accurate estimates of final yield it is usually impossible to judge how closely the forecasts represent true yield. For commercial crops, independent data relating to yield are available through records of arrivals in markets or at factories; but these are ordinarily far too incomplete to provide an effective check on the forecast estimates. For grain crops even this information is lacking. It is frequently argued that the judgment of a skilled and experienced observer regarding average yield cannot be much wrong; but this contention has not been borne out whenever it has been put to a test (Yates, 1936; Irwin, 1938). Yates (1936) has given interesting examples of how forecasts based on a casual inspection of the crop can go badly wrong, and how biased results are produced either by attempting to choose deliberately an average sample or by omitting to follow an objective procedure in sampling. Agreement between different observers is no guarantee that the estimate represents the true average closely. All or majority of them may systematically under or over-estimate it. This bias can be allowed for only if its magnitude and direction can be shown to be fairly constant.

1. Institute of Plant Industry, Indore. 2. Department of Agriculture, Central Provinces.

To be of any real service, both forecasts and estimates of yield must be free from bias and largely free from accidental errors to which all estimates are subject. The efficiency of present methods of crop estimation in this respect needs testing, and the employment of objective procedures for this purpose, if not for replacing present methods altogether, are attracting increasing attention of Government departments concerned everywhere. To secure unbiased estimates, random sampling, which gives every member of the population an equal chance to contribute to the estimate, is the only known method available, and if the sampling is sufficiently extensive, accidental errors can be reduced to any desired level. Estimation of crop yield at harvest falls very properly within the realm of application of this kind of sampling. The problem of forecasting can be dealt with by the same method, though estimation is rendered more difficult here, because it must be based only on measurable characters of the standing crop. The underlying statistical theory is simple and is explained in considerable detail by Cochran (1939) in relation to the present subject.

Recent attempts to determine the yield of wheat in several districts of Great Britain and in a portion of North Dacota State in the United States by harvesting and weighing small samples of the crop in randomly selected fields are described by Cochran (1939) and by King and Jebe (1940) respectively. Due probably to the preliminary nature of the experiment, farms were not selected at random in England, while in North Dacota, a perfectly random selection of fields was considered impracticable and route sampling was adopted as a substitute. The unbiased character of the resulting estimate of yield and the validity of its sampling errors thus become suspect due to these deviations from strict randomization in both the cases; but nonetheless these surveys provide excellent illustrations of the application of the principle to the commercial crop.

In India, difficulties of adopting full randomization, whatever their nature and extent in America and England, are not insuperable, thanks chiefly to the existence of the Government land revenue machinery. A recent noteworthy attempt to employ this procedure in a large-scale agricultural survey is due to Mahalanobis (1939, 1942), who applied it to the problem of estimating the annual area under jute in Bengal, where due to the Permanent Settlement no village staff is available to carry out complete enumeration of area under different crops, a practice which prevails in other provinces. In devising a suitable sampling technique, Mahalanobis is primarily concerned with the evaluation of the relationship between the size of sampling unit and variance between samples, and through field as well as laboratory sampling he devises an empirical function which he terms the variance function analogous to one shown by Fairfield Smith (1938) from uniformity trials. A second similar relationship, the cost function, is further worked out between cost on the one hand and variance and size and density of samples on the other. He proposes to employ the same approach to the problem of estimating yield, though actual results of any such project do not appear to have been published yet.

This method, though interesting mathematically, appears to be of limited utility in connection with the practical problem of sampling for estimation of yield. Exploratory surveys are needed to ascertain the form of the cost and variance functions and to find the numerical values of the statistical constants of these functions. This creates an impression in the minds of administrators and other laymen that complex, time-consuming and expensive research is necessary before practical recommendations can be made about the sampling procedure to be followed. In the jute area survey, preliminary work covering an area of some 2,300 sq. miles, or rather less than the size of one revenue district, involved an expenditure of 155 thousand rupees in a period of three years and an actual survey of all districts growing jute, with a total area of 59,000 sq. miles, cost the same amount of money in one season. The cost in either case was equally divided between field work and statistical analysis. Since the level and variability of crop yields change from season to season and from tract to tract owing to the influence of various environmental factors, it cannot be expected that the variance and cost functions determined from sample data over a restricted area will predict with any great exactness the optimum type of sampling for different tracts and in different seasons (H.O.H., 1941). Estimates of cost based on these functions are not likely to be realized even approximately in practice, when it is considered that the field work on the recommended plan will ultimately be in the hands of the permanently employed revenue and agricultural staff. This condition also implies that a simple and uniform procedure will be highly advantageous. The utility of these functions in providing guidance in sampling for yield estimation will thus not be commensurate with the time and money spent in investigating them.

In fact, if objective methods of crop estimation are to be introduced in India without further delay, we must not place an undue emphasis on the matter of 'technique'. For adopting random selection of sample plots, which is the main feature of such methods, no investigation is required. It is really a question of convincing the Government departments concerned that replacement of the present personal selection of plots by a random selection is both essential and feasible. The optimum intensity and distribution of the sample plots in relation to the particular standard of accuracy desired for the final estimates of yield will, however, have to be determined by experiment. The effect of size and pattern of the sampling unit on the accuracy of the final estimate is of a secondary importance, as the sampling error from this source is only one of the components of error to which the estimate is subject. Commonsense consideration and past experience indicate that the magnitude of this sampling error will be small compared to variation from other causes. Crop estimating surveys can be planned in such a

manner that while the primary objective of estimating yield in the region surveyed is achieved with a known degree of precision, information calculated to increase the efficiency of future surveys both with regard to statistical accuracy and cost is also secured simultaneously. The aim should, of course, be to obtain estimates of yield with the requisite level of accuracy at minimum cost, but progress in this direction will be ensured on the basis of extensive observations made through successive surveys, rather than as a result of special preliminary inquiries in restricted areas. The principle is illustrated below by a brief description of the survey carried out in the year 1942-43 for estimating the yield of cotton in Akola District in Central Provinces. The experiment cost a moderate sum of Rs. 6,000.

The District has an area of 4,092 sq. miles and contains 1,734 villages. It is divided into six administrative subdivisions (*tahsils*). The total area of some 600,000 acres under cotton is distributed among the six *tahsils*, ranging between 69,000 and 137,000 acres per *tahsil*. The number of villages in one *tahsil* ranges from 186 to 348. Cotton is grown in all villages and it is ordinary *Oomra* (*G. arboreum* var. *neglectum* forma *bengalensis*) or its superior strains, Verum 434 and Jarilla. The crop is sown in June on the commencement of the monsoon and is harvested from November to February in five or six pickings. The plan of the experiment was as follows.

In each *tahsil* ten random villages were selected and from a list containing survey numbers of all fields growing cotton in the selected villages in 1942-43, two random fields were selected in each village. In each of these two fields, two plots, measuring 81 ft. × 162 ft. $= {}^{3}/_{10}$ acre were randomly located. Random selection of villages, fields and location of plots was made with the help of Tippet's random numbers (1927). For harvesting, each plot was subdivided into six longitudinal sections of $^{1}/_{20}$ acre size. Yield data of 240 subplots in each *tahsil* or 1,440 plots in the whole District were thus obtained. Besides the *kapas* (seed-cotton) yield of the individual sub-plots, the number of plants, the number of bolls per plant and boll weight were also recorded in small sample units in one of the two experimental fields in each village. The analysis of variance of plot yields (sum of six $^{1}/_{20}$ acre sections) in each *tahsil* pooled over the whole District is shown below.

Analysis of variance of plot yields

Due to	Degrees of freedom	Mean sq.
Villages	54	1595
Fields	60	1150
Plots	120	154

While the magnitude of variation differed in the individual *tahsils*, the relative magnitude of variation from the three sources was more or less similar in most *tahsils*. It should be noted that the mean squares for the fields and villages in the analysis of variance include, besides the real variation due to these items, the sampling variation, since the yields of fields were measured only by sample plot yields, and village yields were based on those of a few randomly selected fields in each village. The real variation between villages is represented by the excess of the mean square for villages over that for fields, and the difference between the mean squares for fields and plots provides an estimate of the real variation due to fields. We thus see that variation from field to field is clearly the most dominant factor affecting the mean yield of a *tahsil*. There is a certain amount of additional variation due to differences between villages; but compared to these two factors, variation between plots in the same field is very small. This relationship between the three variances forms the basis for devising a sampling technique capable of giving yield estimates with a desired degree of accuracy. It is clear from the analysis of variance that the precision of mean yield will depend most on the number of fields sampled and least on the number of plots laid out in each field. Statistically the standard error of the mean yield is related to the variances from the three sources by the formula,

$$Vm = \frac{V}{n} + \frac{F}{nk} + \frac{P}{Nnk},$$

where Vm is the desired variance of square of the standard error of the mean yield, V, F and P the true variances due to villages, fields and plots and n, k and N the number of villages, fields and plots respectively. The estimates of true variances obtained from the present data are

$$\text{Due to villages (V)} = \frac{1595 - 1150}{4} = 111 \text{ per village.}$$

$$\text{Due to fields (F)} = \frac{1150 - 154}{2} = 498 \text{ per field.}$$

$$\text{Due to plots (P)} = 154 \text{ per plot.}$$

By substituting these values in the formula above, the amount of sampling and its optimum distribution for securing a given degree of precision for the mean yield can be determined. To illustrate this on the basis of the present results, the average yield will have a standard error of 5 per cent., *i.e.*, this average will be subject to a maximum chance fluctuation of 10 per cent. on either side of the estimated value, if the following sampling is adopted.

Number of villages (n) required to give 5 per cent. s.e. of mean

No. of fields per village (k)	No. of plots per field (N)	
	1	2
1	194	175
2	111	101
4	70	65
8	49	47
10	45	43

The relative importance of the number of plots per field and of the number of fields per village in determining the number of villages is clearly brought out in this table. The number of villages shown above refers to the whole region whose mean yield is required to be determined with a standard error of 5 per cent. It may be a single district or all cotton-growing districts in the province taken together. In the latter case, the accuracy of mean yields in the individual districts will be naturally much less. Subdivision of the area sampled into smaller units such as *tahsils* of a district is important both for administrative convenience and for securing increased accuracy of the final mean since errors are usually of different magnitudes in the different subdivisions.

The subdivision of plots into $^1/_{20}$ acre sections was intended to provide information on the relative merits of plots of different sizes. Three plot sizes $^1/_{20}$, $^1/_{10}$ and $^3/_{20}$ acre could be compared as in a uniformity trial and the coefficients of variation were found to be 24·1, 20·0 and 18·1 respectively. While there is thus a gradual reduction in error as plot size is increased its magnitude is not appreciable, and when other factors which contribute to the ultimate error of the mean yield are taken into account, it does not appear that the choice of a particular plot size in preference to another is likely to be of any importance in increasing the accuracy of mean yield. Questions of practical convenience will then determine the plot size to be adopted and $^1/_{10}$ acre plots which are at present in use appear quite satisfactory. By replacing $^3/_{10}$ acre plots used in calculating the number of villages required to give a 5 per cent. standard error of the average yield shown in the table above by $^1/_{10}$ acre plots, no material change was observed in the figures.

The cost of field work may be divided into two parts, one being the cost of harvesting which will depend on the total area harvested, *i.e.*, on the total number of plots, and the other would include the cost of travelling and supervision. To save travelling from village to village it is clearly advantageous to reduce the number of villages and increase the number of fields per village. The question of minimum or optimum cost can be examined theoretically; such examination, however, is of little value unless the data on which it is based refer to the routine adopted in practice and not derived from a special inquiry. But with the large-scale adoption of this method for estimating yield relation between cost and statistical accuracy must receive increasing attention with the aim of minimising costs required to attain a prescribed degree of accuracy in the estimates.

The present survey was designed to fulfil the twofold objective of giving an unbiased estimate of the average yield of cotton in Akola District in the year 1942–43 and providing information essential for increasing the statistical efficiency of future surveys. The average yield per acre for the whole district was estimated from the survey at 136 lbs. of seed cotton and this had a standard error of 6·3 per cent. Taking into account the acreage under cotton, the estimated total production of cotton in the district came to 72,652 bales (1 bale = 392 lbs. of lint). It is important to note that the official estimate was over 41 per cent. higher than this figure and well outside the range of chance fluctuation of the experimental estimate. There can be little doubt that the former was a serious over-estimate of the production. On the other hand, it is interesting to record that the cultivators whose fields were sampled in the survey under-estimated their yield by 26 per cent.! With the help of the technical information now available, it is possible to raise the accuracy of the yield estimates, at the same time simplifying the sampling procedure. For instance, without any alteration in the number of villages or the total area harvested, we may introduce the modification that instead of having two plots in each field and two fields per village only one plot needed be located in each field and the number of fields per village increased to four. This is expected to reduce the standard error considerably. A further simplification would result from adopting a uniform plot size of $^1/_{10}$ acre and not subdividing it into sections for harvesting. The yield survey for Akola in 1943-44 was planned on these lines and a more accurate estimate of yield is anticipated without any increase in cost. For verification of the technical results, the survey was also repeated with the previous design in another district.

Forecasting of yield from the standing crop before it is harvested is of considerable importance to the trade and such forecasts based on the inspection of the crop are issued officially. The present survey provides some useful information on the accuracy of forecasts and the means of improving them. The forecast estimates of yield of the fields selected for the survey were made by the inspector in charge of the survey who was accustomed to making such estimates by eye judgment. He over-estimated average yield by 31 per cent.; but apart from this bias his estimates showed a fairly high correlation ($r = \cdot 8$) with actual yields. Corrected for bias, these forecasts would thus provide tolerable estimates of probable yield; but it cannot be assumed that either the magnitude or the direction of bias would remain constant in future seasons or in other districts. Experience elsewhere indicates that the bias changes in an unpredictable manner and it consequently appears impossible to prescribe a correction in advance. More reliable forecasts of yield should result from the use of objective methods such as quantitative observations on the yield constituents of the standing crop. This is particularly feasible for cotton where it is easy to count the number of plants in unit area, the number of mature bolls per plant and, when picking commences, the weight of seed cotton per boll. These observations were made in the present survey but it is not intended to discuss them here beyond mentioning that they did not, contrary to expectation, show a closer association with actual yield than eye inspection and frequently the latter proved superior. The

most probable explanation is provided by the fact that the observers were instructed to count those bolls which in their judgment would open during the harvesting season and contribute to yield; but this led to considerable variation in practice, some observers including in their count quite small bolls, others excluding considerably developed ones. More precise instructions are obviously needed on the type of bolls to be counted. The investigation is being continued.

This brief account of the cotton survey is presented to illustrate the application of the random sampling method to commercial crops as the most reliable means of estimating their yield. It is satisfactory to note that rapid progress is being made in the adoption of this method and sufficient data will soon be available for a more detailed discussion of its various aspects. Besides the extension of the cotton survey to two districts in Central Provinces in the past season, an investigation on similar lines was projected in two districts, one in the Central Provinces and the other in the Punjab, for estimating the yield of wheat. Two large-scale surveys embracing the whole provincial area under wheat were also carried out in the Punjab and the United Provinces last year. Immediate expansion of such yield surveys on a still wider scale can be recommended without hesitation; because the design described above has the advantage that besides giving unbiased estimate of yield for the area surveyed, it also provides data which form the basis for introducing such modifications in the plan of future surveys as are calculated to improve their efficiency. This improvement can go on steadily without in any way affecting the value of previous results. To achieve this end, it is most important, however, that the procedure adopted and the statistical results obtained should be under continuous review by a competent statistician charged with this work and not merely called in to give advice when difficulties arise. At present a number of crop cutting experiments are carried out by the staff of the revenue and agricultural departments and this personnel may be organised and utilized for field work under the new plan.

Not only for estimating yield, but for a variety of other investigations such as the effect of meteorological factors on crop, the spread of diseases and pests, spread of improved varieties of seed and other agricultural improvements, social and economic inquiries among the rural population, random sampling surveys on the pattern described here provide the most satisfactory means. A beginning in this direction is urgently necessary because past surveys of this kind have been frequently and justifiably criticized on the question of their respective character and the accuracy of their results.

The present survey was financed by the Indian Central Cotton Committee.

1. Cochan, W. G., *J. Amer. Stat. Assn.*, 1939, **34**, 492–510. 2. Fairfield Smith, H., *J. Agri. Sci.*, 1938, **28**, 1–23. 3. H. O. H., *Plant Breeding Abstracts*, 1941, **9**, 71–72. 4. Irwin, J. O., *Suppl. J. Roy. Stat. Soc.*, 1938, **5**, 1–45. 5. King, A. J., and Jebe, E. H., *Research Bull. 273*, Agri. Expt. Station, Iowa, U.S.A., 1940, 624–49. 6. Mahalanobis, P. C., *Sankhya*, 1939, **4**, 511–31. 7. —, *Presidential Address*, Indian Science Congress, Baroda. Section of Mathematics and Statistics, 1942. 8. Tippet, L. H. C., "Random Sampling Numbers," *Tracts for Computors*, No. XV, 1927, Cambridge University Press, London. 9. Yates, F., Manchester Statistical Society, 1936, pp. 1–26.