YRB 54-02

A Determination of Optimum Plot Size for Estimating Mean Number of Bolls

by

Daniel E. Sands

June, 1954

Table of Contents

I. Background and Statement of Problem

1.1 Cotton Crop Estimates (In General)

The Crop Reporting Board, U. S. Department of Agriculture each year, at certain intervals fixed by law, makes three estimates of cotton acreage and five estimates of cotton production. The first estimate, which reports the number of acres of cotton in cultivation as of July 1, appears about July 8. This figure is essentially unmodified until December 1, the only revision taking place on September 1 when reduction in acreage due to abandonment is estimated. On December 1 the Board issues a revised acreage estimate on the basis of new acreage reports obtained from farmers and on any other new information that may have become available.

Production estimates are made monthly from August to December, with the early reports relying principally on information received from farmers and other crop reporters. The August estimate of production is dependent upon the July 1 acreage estimate corrected by the 10 year average abandonment for each state. In the September schedule the reporters are asked for abandonment figures, and, as a result, the September production is based on an acreage figure corrected by estimated actual abandonment that year. The October and November production estimates make use of the September acreage figures. From October onward figures of bales ginned and estimated number of bales to be ginned are available from the ginners, as required by law. These figures serve as an additional source of information. Finally, in its December schedule, the Board asks reporters for acreage abandoned since July 1 and also requests a new figure for acres in cultivation from which it revises its July 1 estimate.

Table I - Reports Issued by the Crop Reporting Board for Cotton

July 1	- Estimate of acres in cultivation
August 1	- First production estimate
September 1	- Modification, due to abandonment, of July acreage estimate
	- Second production estimate
October 1	- Third production estimate
November 1	- Fourth production estimate
December 1	- Revision of estimate of acres in cultivation on July 1.
	- Fifth production estimate

1.2 The 1951 Cotton Crop Estimate and its Effect

The estimates for the year 1951 are of special interest because their lack of accuracy led to a series of inquiries by a congressional subcommittee into the estimating procedure. One consequence of these inquiries is this thesis. Pertinent 1951 data are summarized in Table II. 1/

The first acreage estimate in 1951 was 29.5 million acres. As a result of the schedules mailed out for December, the acreage estimate was reduced by 1.5 million. The revised figure, 28 million acres, thus was the figure that actually should have been the estimate as of July 1. However, the first estimate was modified on September 1, so that the figure 28.5 million acres was the basis for production estimates, a value too large by approximately .5 million acres. In previous years the difference between the July and December acreage estimates had been only 1.3%, but in 1951 it was 5.4%. Partly as a result of this error, the earlier production figures were much too high. The early season yield estimates were also too high.

1/ Agricultural subcommittee report. 1952. Crop estimating and reporting service of the Department of Agriculture. Report and recommendations of a special subcommittee of the Committee on Agriculture of the House of Representatives, 82nd Congress second session. U. S. Government Printing Office, Washington, D. C.

In August, the first production estimate forecast a crop of 17.27 million bales, 7.30 million more than the previous year and 3.30 million bales above the 10 year average. The estimate was raised to 17.29 million bales in September, and then reduced to 16.93 million bales in October, a decrease of 360,000. In November 1.16 million bales were subtracted from the October figure to bring the estimate down to 15.77 million bales. Finally, the December estimate was 15.29. This last figure was not too far from the 15.13 million bales given by the May, 1952, final estimate of production.

Since the market price of cotton is delicately attuned to the Crop Reporting Board estimates, the price paid to the farmer for his cotton changed markedly after each monthly estimate. Following the July estimate of 29.5 million acres the market price dropped from the 45 cents a pound figure brought by the short 1950 crop to 40.82 at the end of the second week in July, and by the end of July was down to 36.21, which was 9.04 cents below the figure paid in the last day in June. The average price was down to 35.08 cents on August 7, the day before the first production estimate, and the end of August the price was steady at 34.32. The slightly higher September estimate caused little price change, but some doubts about the accuracy of the estimates pushed the price up to 36.31 at the end of September. Following the reduced October forecast the price moved up to 38.20 at the end of that month. When the November report came out with its reduction of 1.16 million bales, the price jumped to 41.45 cents immediately and reached 43.16 by November 30. After the comparatively minor change in the December estimate, the price stabilized at around 42 cents per pound.

The result of all this price fluctuation was that cotton producers who sold their crops earlier in the season, that is, between July and October, were deprived of approximately $25.00 per 500 pound bale of ginned cotton. It has been

estimated that during this period farmers sold about 5 million bales of cotton, thus losing approximately $125 million as a direct result of the Crop Reporting Board's overestimate.

Tale II - Crop Estimates and Reported Prices for Cotton, 1951, 1/ by Release Dates.

		Estimate		Price (cents per pound)
		acres	(millions) bales	
July	9	29.5		43.69
	13			40.82
	27			36.82
Aug.	8		17.27	34.82
	17			34.96
	31			34.32
Sept.	10	28.5	17.29	34.22
	21			35.29
Oct.	8		16.93	36.87
	19			36.63
	26			37.53
Nov.	8		15.77	41.45
	16			41.67
	30			43.16
Dec.	10		15.29	42.85
	21			41.78

1/ Agricultural subcommittee report. 1952. Crop estimating and reporting service of the Department of Agriculture, Report and recommendations of a special subcommittee of the committee on Agriculture of the House of Representatives, 82nd Congress, second session. U. S. Government Printing Office, Washington, D. C.

1.3 The Problem

1.3.1 Three reasons advanced for the failure to make an accurate estimate were

(a) the schedules contained vague or complicated questions,

(b) the crop reporters were not a wisely chosen group,

(c) some form of objective measurement should be used.

To overcome the schedule problem, (a), the congressional subcommittee recommended that unambiguous questions of a factual instead of a judgment nature be asked, a reporter be questioned only about his own operations, and questions be phrased so that answers would be descriptive instead of in percentages. In reference to crop reporters, (b), the subcommittee suggested that lists be kept up to date, non-respondents and those who do not supply information with reasonable accuracy be eliminated, local officials select competent and interested reporters, and alternative methods such as telephone contact be used to secure information from reporters. With respect to (c), the subcommittee believed that objective measurements such as cotton boll counts, crop metering, and ginning reports should be used.

1.3.2 A production forecast may be considered to be made up of the following factors.

(1) An estimate of the total acreage.

(2) Measurement of one or more characters related to eventual yield.

(3) Extrapolation to a yield per acre forecast.

(4) Multiplication of the acreage estimate by the yield per acre forecast.

The discussion will be restricted to factor (2). The character measured was large bolls, and counts of these were made on randomly selected samples of small areas in cotton fields. It must be determined what size these small areas are to be, how many are to be sampled in each field and how many fields are to be included, keeping in mind the fact that the results must have the least variability

for the amount of money that can be spent. The part of factor (2) that concerns us here is the size of the sample area or plot.

An estimate of plot size involves two concepts, precision and accuracy. By precision is meant the variability around the average value obtained in repeated sampling. Accuracy may include another component, bias, which may arise, for example, from including border plants just outside the plot.

Specifically, the purpose of this thesis is to discover the optimum size of sample plot to be used for making counts of large bolls. Here, optimum means the size of plot which will give the least variability (most precision) in the boll count values for a given expenditure of funds.

II. Review of Literature

2.1 Measurement Errors With Small Plots

Sukhatme (1947) gives the results of three investigations carried out in India for the purpose of comparing the efficiency of small size plots with those of the order of 1/80th of an acre used under the existing official procedure in India.

The plot sizes used in India are comparatively large, being 1/100, 1/40, 1/10, and 1/160 of an acre as compared to 1/4000 and 1/5000 used in England by Cochran (1939), who described a plot size of 1/4 meter by 6 rows in a field of grain. The investigations in India were carried out in 1944-45 to test the efficiency of small plots of varying shape.

Results were that the estimate for the average yield decreased as the plot size increased, and the rate of decrease diminished with increase in the plot size, which suggests that when the plot is sufficiently large, the estimate attains a stable value, and that when the plot is less than 30 sq. ft. in size there is a serious overestimation. This bias was attributed to the tendency to include

border plants inside the plot area, which in the case of small areas has a great inflating effect.

The conclusions were that small plots, even when observed with extreme care, cannot be depended upon to give unbiased yield estimates. Large plots, of the order of 1/80th of an acre appear to be free from this source of bias.

Yates (1935) reported that when sampling in a field was not carried out with complete objectivity the tendency of the observer was to select a better than average area of the field. First, lengths of one meter were selected along rows of wheat, out of which the observer was to select one quarter of a meter. Almost invariably he would choose the most dense quarter of a meter. This method of sampling led to estimates 13 to 35% too high.

In a study of plot size for rice production estimation, Mahalanobis (1946) found the following positive biases in comparison to a plot of 576 sq. ft.; 9 sq. ft., 13.3%; 36 sq. ft., 12.1%; 144 sq. ft., 1.2%.

Mahalanobis (1945-1946) found further that when experienced observers harvested sample plots under the supervision of trained statisticians, a previous upward bias of 14.9% (when compared to a circular area of 100 sq. ft.) decreased to almost zero. The conclusions were that if, under proper supervision, random points in the field were chosen objectively and if plants just outside the plot borders were excluded, accurate estimates could be obtained using small plots.

2.2 Reliability of Small Plots

Johnson (1943) in working out a method for improving sampling procedures for tree nursery inventories, made a study of optimum plot size which is of particular interest because the actual population values were known. Complete counts were made of seedlings in twenty seed beds and transplant beds each of which was approximately 400 feet long and 52 inches wide. In some beds the seeds had been

sown in rows and in others it had been sown broadcast. The sampling units used in the beds planted in rows were one, two and three feet units of single rows and one, two and three foot widths of the entire bed (across all six rows). In the beds in which trees were not planted in rows a 6 inch sampling frame was used which gave three sampling unit sizes, 1/2, 1, and 1 1/2 foot units of the bed width.

The conclusions reached were that the most efficient units for estimating the total tree stock in hardwood seed beds planted in rows were 1 ft. row, 2 ft. row, and 1 ft. bed width units depending on the hardwood species. For coniferous seedbed stock the best units were the 1 and 2 ft. row units. For coniferous stock sown broadcast the best unit was the 1/2 foot bed width unit, while for the coniferous transplant beds the one foot bed width unit was the most efficient.

Hameed (1953), working with corn in Iowa, considered eight plot sizes for the purpose of choosing one of them for use in the estimation of corn yield. The plots were 50', 100', 150' and 200' of both one and two rows. In each of 18 fields located on a total of four farms, plots of 200' x 2 rows were selected by a system involving use of a table of random numbers. Each of these plots was split into eight 50' x 1 row plots and a 1/5 systematic sample of ears was taken in each of the small plots. The variable considered was weight of the grain produced.

The sampling estimates of the variances were obtained by analysis of variance and the relative efficiencies calculated. It was found that the relative efficiency decreased with increasing plot size and that the relative statistical efficiency of the smallest plot was greatest.

A cost function using the estimated amount of time needed to perform various items of work was set up. (No records of time or cost were maintained). This

included the time to choose, mark off, enumerate the plots in a field plus the time necessary to travel between plots.

The net relative efficiencies were then calculated, and it was found that the plot size 50' x 2 rows was the most efficient among the eight sizes considered when cost was taken into account. It was estimated by a mathematical technique that the optimum is between 20' and 25' of two rows.

Smith (1938) devised an empirical law for the purpose of measuring soil heterogeneity quantitatively. Using results of a uniformity trial with wheat he observed that between plot variability decreased with increasing plot size, as had been found in many other experiments. The relationship was linear when plotted on log log paper, and he concluded that the equation

$$V_x = \frac{V_1}{x^{b'}} \qquad (1)$$

or

$$\log V_x = \log V_1 - b' \log x \qquad (2)$$

expressed the relationship between variability and plot size, where

V_x is the variance of yield on a per unit basis among plots of x units

V_1 is the variance of single unit plots

x is the number of units in a plot

b' is a linear regression coefficient expressing the regression of the log of the variance per plot on the log of the number of units (x) in a plot, the observations weighted according to their degrees of freedom.

This empirical relationship was applied to results from thirty-nine experiments and most of these relationships were satisfactory.

The regression coefficient, b', may be regarded as reflecting the soil and plant heterogeneity in a field. A low degree of correlation (high degree of heterogeneity) between adjacent small units in the field, as expressed by a small V_x, will result in a high value of b'. If the small units were completely uncorrelated $b' = 1$ and equation (1) becomes

$$V_x = \frac{V_1}{x}$$

the variance of the mean.

The above discussion holds for a field with a fixed number of plots n, but for the relationship to be of more value it should hold for a field of any size. Therefore, an infinitely large field is postulated of which an observed field is a single block, and an adjustment is applied to b' so that a modified regression coefficient b is applicable to an "infinite" field. Then, the variance of x units within blocks of m plots is obtained by analysis of variance.

Using the infinite model and inserting a cost function leads to an expression of optimum size of plot. When K_1 is the cost of a replication and K_2 the cost of a plot, cost per unit of information is at a minimum when

$$x = \frac{b\,K_1}{(1-b)\,K_2}$$

A method of using experimental results in the place of uniformity trials for the purpose of estimating optimum plot size is discussed by Koch and Rigney (1951). Variances of units of several sizes were reconstructed from the variance components obtained from the analysis of split-plot and incomplete block designs. After computing a regression coefficient, b, the empirical technique of Smith (1938) was used to estimate optimum plot size to be used

in future experimental work involving the effects of various treatments.

Peanut plot uniformity data were interpreted by Robinson, Rigney and Harvey (1948) for the purpose of finding the optimum size and shape of plot. Two methods were used. First the coefficients of variability, i.e., the standard deviations divided by the means, of various plot sizes were computed. These were scaled by a cost function under the assumption that a certain part of the cost was proportional to the number of replications (K_1) and the remainder of the cost was proportional to plot size (K_2). As K_2 increased the relative cost for a given number of 12 1/2 foot units per plot increased.

The second method was the regression method of Smith (1938). The necessary regression coefficients, b, were computed and the optimum plot size was obtained for soils that differed in heterogeneity, as expressed by the b values. Costs of plots and units were introduced and results presented in a graph under the same assumption as above. The graph shows optimum plot sizes for various combinations of K_1 and K_2.

III. The Data 1/

3.1 Method of collection.

It has been felt for some time that improvements in crop forecasting techniques were necessary, and under the added stimulus of the inaccurate 1951 cotton estimates some work was undertaken in the "investigation of methods for improving accuracy of cotton production forecasts." A project was set up with the cooperating agencies being the Bureau of Agricultural Economics of the U. S. Department of Agriculture, The North Carolina Cooperative Crop Reporting Service, and the Institute of Statistics at North Carolina State College. This project had two

1/ The original data and the tabulation sheets are on file at the North Carolina State College, Department of Experimental Statistics.

distinct phases.

We shall discuss only the second phase which consisted of a mail and personal interview survey in Johnston and Harnett counties, North Carolina, during the 1953 cotton growing season. Among the objectives of the survey was the estimation of the optimum size of sample plot to use in making counts of large bolls, i.e., bolls as large in diameter as a twenty-five cent piece. For this purpose data were secured by interviewers in August and in September, 1953, during a series of personal interviews made during that summer with a randomly selected group of cotton farmers.

During the August 1 personal interview of cotton growers, the interviewers were requested to carry out some additional measurements in eighty fields after the main schedule had been completed. These additional measurements were carried out as indicated below.

Each interviewer was given a list of cotton growers to be visited. (See Figure 1). After the interview had been completed, the interviewer would proceed to the northwest corner (A)[1/] of the second field nearest the farm headquarters. At this place he would note the time and then go twenty rows along the edge of the field (B), twenty paces down the row (C) and drive a long stake in the row to the right. This was the initial point of the sample section, which was to consist of a twenty-five foot length of two adjacent rows (25'D). This in turn, was to be subdivided by the use of small stakes into ten smaller sections, each consisting of a five foot length of a single row (5'S). But before doing this, the interviewer was to consult the written instructions made up for each field. These told him first to mark with small stakes a specific number of five foot lengths within the 25'D sample section and to record the number of large bolls

1/ The letters in parentheses refer to specific locations on Figure 1.

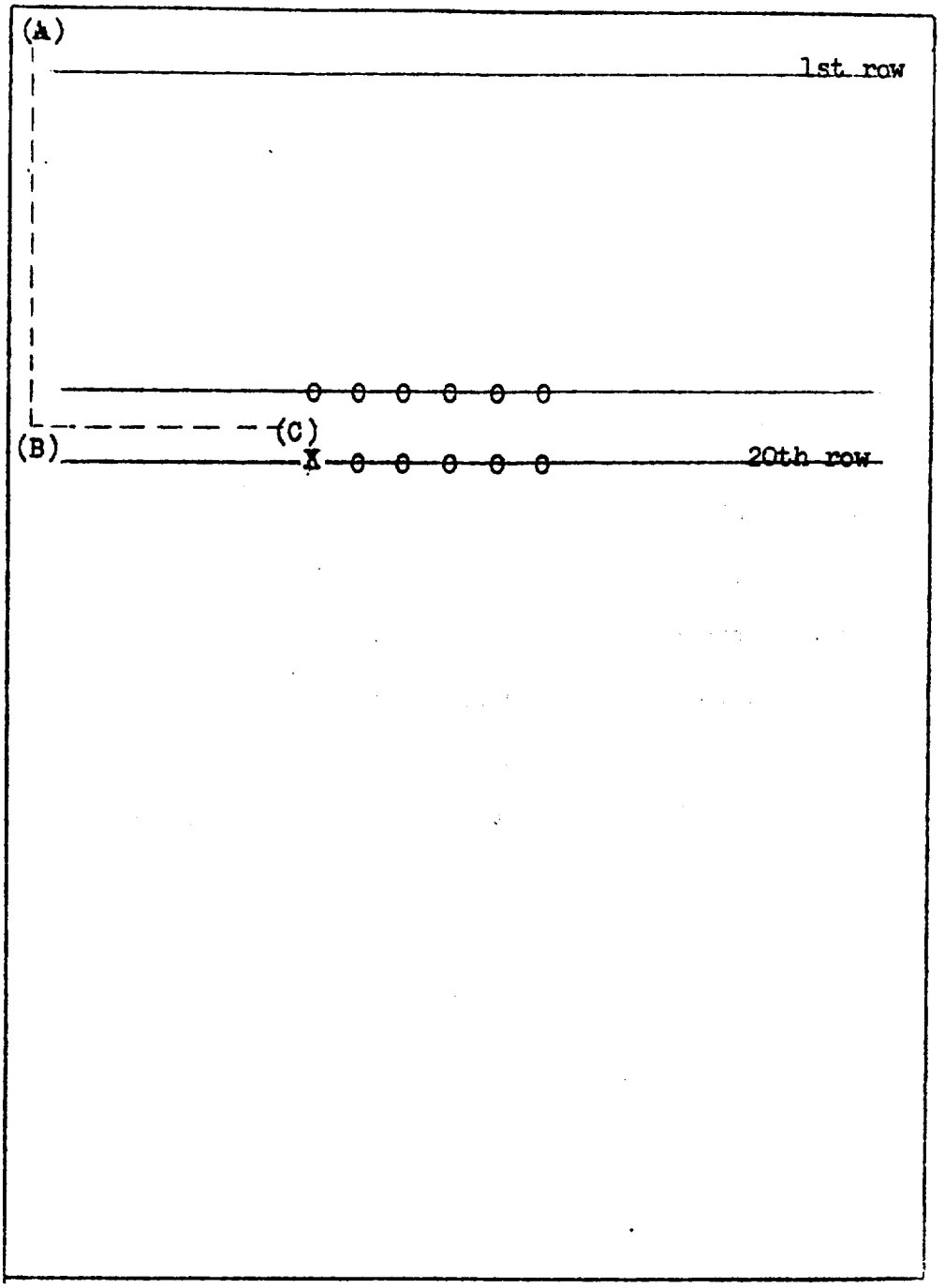


Figure 1. Diagram of Field Illustrating Method of Selecting Sample Sections.

and the amount of time in minutes and seconds necessary for him to locate the large stake and count the number of large bolls in each five foot section of the group. Thus, depending on the field he was in, the interviewer would stake off within the 25'D section a portion consisting of 5, 10, 15, 20 or 25 feet of one of the rows, or 5, 10, 15, 20 or 25 feet of both of the rows. On a form provided for each field he would enter the total time and the number of large bolls in each 5'S he had just counted. The times needed to locate and make large boll counts on the ten possible sizes of plots within the 25'D sample section play an important role in the analysis, since they represent the cost of labor involved in making the counts.

Following this the number of large bolls were counted on the remainder, if any, of the 5'S sections within the 25'D sample section. The counts were recorded and the total time elapsing since he left the northwest corner of the field was entered on the sheet. The short stakes were pulled up and the long stake was left in the field.

The counts secured by the method outlined above and counts of large bolls made on a randomly selected ten feet of row as a part of the schedules were used in the August 1 computations.

During the September 1 personal interview, counts of large unopen bolls ("large" meaning at least the diameter of a twenty-five cent piece) were made on the same ten foot section of single row used in August. Again additional information was collected but this time portions of two 25'D sampling units from each of forty fields were selected on which to take measurements in order to obtain an estimate of the variability between units of the same size in the fields. These additional measurements were completed by the interviewers according to the instructions which follow.

The long stake driven in August was located and counts of large unopen bolls were made in predetermined parts of the same 25'D sample section used in August, and the time recorded. The interviewer then located another 25'D sample section in the same field by crossing over twenty rows from the original large stake, proceeding twenty paces toward the center of the field, and driving a long stake to the right. The time since leaving the first large stake was recorded. Five foot sections of single row were staked off and counts of large unopen bolls were made in these indicated on a diagram. The total time since leaving the first large stake was recorded.

As a result of the work done in August, the information made available for plot size estimates included

a) large boll counts on one 25'D section on each of eighty fields, each of the ten 5'S sections being counted separately.

b) the time needed for locating and making counts on a prescribed portion of the 25'D sample section.

c) a count of large bolls on a 10'S section in a randomly selected part of the field, done as part of the schedule.

In September large unopen boll counts were made on forty fields. From each field there was obtained

a) the count on a portion of the 25'D area staked off in August,

b) time needed to locate and count a portion of a second 25'D area

c) the count on a portion of the 25'D area staked off in September, and the time needed to make the count.

d) counts of large unopen bolls in a 10'S section made as part of the schedule.

3.2 Limitations

The data which made this study possible were gathered under conditions which help explain the erratic behavior of the results and the small number of degrees of freedom associated with some of the variance components, as will be noticeable in Chapter IV. These conditions fall into two groups, environmental and administrative.

The cotton survey in Johnston and Harnett Counties was conducted during a period of severe drought in some sections of the area. In other sections there were occasional rains, but to a large extent they did not come when they were needed most. The lack of moisture coupled with the fact that most of the farmers in the area are tobacco farmers and consequently have small cotton acreages, caused many of the farmers to have little interest in their cotton crops. The area was chosen because its nearness to Raleigh helped keep down the travel and field expenses, and also the personnel involved in the study had other obligations to fulfill in Raleigh.

The critical shortcoming, however, was in the procedure by which the data were obtained. During the August survey only one 25'D sample section was taken in each field. This prevented any estimate of the variability between the larger sized units within fields from being made easily. More specifically, each of fifty out of the eighty fields visited in August contained one 25'D sample section and one 10'S section on which counts of large bolls were made as part of the interview schedule. Thus, the required estimates of the variances between the larger plot sizes within fields, say 10'D or 25'S, could not be obtained by the usual and more familiar computing procedures.

In September a similar situation was encountered. It was intended that there be two 25'D sample plots in each of forty fields, but after making the boll counts and recording the time and the counts according to the directions for each field,

the interviewers did not make boll counts on the remainders of the 25'D sample sections. As a result the data in each field were not complete. For example, one field might contain one 15'D section and one 10'S section, bringing up the same difficulty encountered in August. However, some results involving the various plot sizes were obtained for the September data, but in the case of the larger plot sizes there were too few degrees of freedom.

Another difficulty was superimposed on the existing situation. In August, when nearly all of the bolls in the area are not yet open, the variable measured was large bolls. This included all bolls at least the diameter of a twenty-five cent piece. In September, when a large number of the bolls have opened, the variable measured was large unopen bolls. Hence, in August all the large bolls were counted, while in September only a part of them were counted, the large open bolls not being included in the counts. This fact made direct comparisons between August and September results difficult to interpret.

IV. Analysis of Data and Results

The statistical analysis was carried out in four steps. To find the optimum plot size from among the ten possible sizes for which data are available, the steps are to find for each plot size

(1) the variance between units of the same size within fields, σ_B^2 ,

(2) the relative statistical efficiency, RSE,

(3) the relative cost, RC,

(4) the net relative efficiency, NRE.

That plot size which has the largest net relative efficiency is defined to be the optimum, i.e., the mean boll count for plots of that particular size will have the least variability with a given expenditure of funds.

4.1 Methods

4.1.1 Mean Squares and Variance Components

4.1.1.1 Analysis of Variance

The farms and fields were randomly selected and for purposes of analysis, the 25'D and 5'S sample sections within fields were assumed to be randomly selected. The number of 25'D sample sections within fields is considered as being infinite, as is the number of 5'S sections in each 25'D, and the ten 5'S sections actually chosen are assumed to be a random sample from these. Now, considering 25'D as the sampling unit and any smaller area within the 25'D as the sub-unit, the observation x_{ij} from the jth sub-unit of the ith unit may be explained by the identity

$$x_{ij} = \mu + b_i + w_{ij} \qquad (1)$$

where μ is the population mean on a sub-unit basis

b_i is the deviation of the mean of the ith unit from the population mean, μ.

w_{ij} is the deviation of the jth sub-unit in the ith unit from the unit mean.

The means of b_i and w_{ij} are defined to be zero and their variances are σ_b^2 and σ_w^2. In addition, all b_i and w_{ij} are assumed to be independent. All the foregoing statements in this section may be summarized by saying that the assumptions underlying the analysis of variance are assumed to be satisfied.

The analyses of variance used were of two forms.

(i)

Source	d.f.	E(MS)
Between fields	(f-1)	
Between units in fields	f(m-1)	$\sigma_w^2 + n\sigma_b^2$
Between sub-units in units	fm(n-1)	σ_w^2

where $\underline{f}$ is the number of fields

$\underline{m}$ is the number of units in a field

$\underline{n}$ is the number of sub-units in a unit

d.f. is the number of degrees of freedom

E(MS) is the average over all possible sample mean squares.

In subsequent discussions E is the expectation or average over all possible samples.

(ii)

Source	d.f.	E(MS)
Between fields	(f-1)	
Between sub-units in fields	f(m-1)	σ_w^2,

In (i) the variance between units in fields is given by $\sigma_w^2 + n\sigma_b^2$ and the variance between sub-units in fields is given by $\sigma_w^2 + \sigma_b^2$. The latter variance is shown by using (1), and considering the assumptions underlying the analysis of variance

$$E(w_{ij}) = 0 \quad \text{and } E(w_{ij})^2 = \sigma_w^2$$

$$E(b_i) = 0 \quad \text{and } E(b_i)^2 = \sigma_b^2$$

it follows that

$$E(x_{ij}) = \mu$$

Then,

$$V(x_{ij}) = E(x_{ij} - \mu)^2$$

where $V(x_{ij})$ is the variance of a randomly selected sub-unit in a field.

By substitution using (1)

$$V(x_{ij}) = E(b_i + w_{ij})^2$$

$$= E(b_i)^2 + 2E(b_i w_{ij}) + E(w_{ij})^2$$

Since independence is a sufficient condition for the covariance to equal zero, the middle term drops out and it follows that the variance between sub-units within fields is

$$V(x_{ij}) = \sigma_b^2 + \sigma_w^2 \qquad (2)$$

In (ii) the variance between sub-units in fields is simply $\sigma_{w'}^2$.

An important fact that must be borne in mind throughout this entire study is that the mean squares and variance components used were placed on a sub-unit basis. The basic sub-unit was 5'S. When an analysis of variance of the type (1) was computed and 5'S was the sub-unit, the variances were automatically on a 5'S basis. When any other plot size basis was used in the analysis, variances were divided by an appropriate integer in order to place them on a 5'S basis. Divisors used are shown in Tables VI, VII, VIII. The reason for this is to make the variances directly comparable with each other, under the assumption that the same total area of the population is sampled in both cases.

4.1.1.2 Alternative Method

Due to the limitations set forth in 3.2 an alternative method for estimating the variances between units of the same size in fields had to be devised. This was done because in many cases the data were not sufficient to enable a precise estimate of the variance to be made using the analysis of variance technique. The method is as follows.

Suppose we randomly select two units, 1 and 2, of the same size from a large number of units in a field, and suppose their total boll counts are known. In this case the variance between units of this size in fields, σ_B^2, is estimated by one-half the square of the difference between the two boll counts. Symbolically for a single field, this is shown to be an unbiased estimate as follows.

$$E\frac{(x_1 - x_2)^2}{2} = \frac{E(x_1)^2 + E(x_2)^2 - 2E(x_1x_2)}{2} = \sigma_B^2 \qquad (3)$$

Suppose now that the total boll count on unit 1 was known and the boll count on only a fractional part of unit 2 was known. To estimate the variance of between, say 20'S in fields, the data available in each field may consist of a 20'S sample section and a 10'S sample section. In order to estimate $\sigma^2_{20'S}$, the 10'S sample section is assumed to be a randomly selected subsample from a 20'S unit (not completely measured), which introduces a second stage of sampling. The boll count on this 10'S is multiplied by a constant, $k = 2$, in order to estimate x_2. The estimate of σ_B^2 in this situation is obtained indirectly. First the quantity

$$\frac{(x_1 - \hat{x}_2)^2}{2}$$

is computed for each field, where $\hat{x}_2$ is an estimate of the boll count on a unit the same size as unit 1, and is obtained by multiplying a smaller sized unit by the proper constant, k. $\hat{x}_2 = kx_2$ or x_2 is estimated by kx_2. Also, $E\hat{x}_2 = x_2$.

This quantity estimates

$$\sigma_B^2 + \frac{k^2}{2}\sigma^2_{2'},$$

where $\sigma^2_{2'}$ is the variance between units the size of 2' within units the size of 1.

The derivation of this result is as follows.

Let E_1 be the expectation over the first stage of sampling and E_2 be the expectation over the second stage of sampling. Then (3) becomes

$$E_1E_2 \frac{(x_1 - \hat{x}_2)^2}{2} \tag{4}$$

Expanding (4) gives

$$E_1E_2 \frac{x_1^2 - 2x_1\hat{x}_2 + x_2^2}{2}$$

$$= \frac{E_1(x_1)^2 - 2E_1\left[x_1E_x(\hat{x}_x)\right] + E_1E_2(\hat{x}_2^2)}{2}$$

$$= \frac{(\sigma_B^2 + \mu^2) - 2E_1\left[x_1x_2\right] + E_1E_2k^2x_2^2}{2}$$

$$= \frac{(\sigma_B^2 + \mu^2) - 2\mu^2 + k^2E_1\left[E_2(x_2^2)\right]}{2}$$

$$= \frac{\sigma_B^2 - \mu^2 + k^2E_1\left[\sigma_2^2 + \left\{E_2(\frac{x_2}{k})\right\}^2\right]}{2}$$

$$= \frac{\sigma_B^2 - \mu^2 + k^2E_1\sigma_2^2 + (\sigma_B^2 + \mu^2)}{2}$$

$$= \frac{2\sigma_B^2 + k^2E_1\sigma_2^2}{2}$$

Assuming σ^2_2 is the same for all units in fields, $E_1\sigma_2^2 = \sigma_2^2$.

Therefore

$$E_1E_2\frac{(x_1 - \hat{x}_2)^2}{2} = \sigma_B^2 + \frac{k^2}{2}\sigma_2^2$$

For n fields an estimate of σ_B^2 can be obtained by using the expression

$$\sigma_B^2 = \frac{\sum^{n}(x_1 - kx_{2'})^2}{2n} - \frac{k^2}{2}\ \sigma_{2'}^2 \qquad (5)$$

The quantity $\sigma_{2'}^2$ was estimated from the observations available in unit 1 on two units the size of 2'. Under the assumption that the units the size of 2' available from unit 2 were selected at random from that unit, $\sigma_{2'}^2$ is more than likely an overestimate. The reason for this can be explained by the use of an example.

Let unit 2 be of size 20'S and unit 2' be of size 10'S. The two 10'S units are assumed to be a random sample from the 20'S. Since the 20'S is considered to contain a large number of possible 10'S units, most of the 10'S units overlap. Then, two of them picked at random would tend to have similar boll counts, and the the difference between the two values would be small. As a result $\sigma_{10'S}^2$ would tend to be small. However, the figures actually used in the computations are the boll count values for the two non-overlapping 10'S units in the 20'S unit. These two values would tend to have a greater difference than any other two, so a larger value of $\hat{\sigma}^2_{10'S}$ results.

The important consequence of the overestimate of $\hat{\sigma}^2_{10'S}$ is that $\hat{\sigma}^2_{20'S}$ is underestimated, by equation (5). The seriousness of the underestimate is not known, but it is not considered sufficient to invalidate the method.

4.1.2 Relative Statistical Efficiency

If the variance components associated with the various plot sizes have been put on a sub-unit basis (5'S), as discussed in 4.1.1.1 and if the variance of the 5'S sub-unit is

$$V(5'S) = \sigma_{5'S}^2$$

and the variance of some other plot size, say B, is

$$V(B) = \sigma_B^2$$

then the relative statistical efficiency of B to 5'S is defined as

$$RSE = \frac{\sigma_{5'S}^2}{\sigma_B^2} \times 100 \ . \qquad (6)$$

In this study the plot of size 5'S will be taken as the "standard unit" equal to unity, and the relative sizes of the other plots will be expressed by an integer from 2 to 10.

The RSE may be explained in a different and perhaps more meaningful manner. The variance of the boll counts for each plot size is a measure of the spread of the values around the mean, and, intuitively, the more stable the boll count values the narrower the spread. In turn, the narrower the spread the more the trust that can be put in the estimate of the true mean count per plot. This idea has been used to define a quantity known as information, I, where

$$I = \frac{1}{V}$$

Then the ratio of the information offered by plots of size B to plots of size 5'S is a measure of the relative statistical efficiency.

$$RSE \propto \frac{I_B}{I_{5'S}} = \frac{\frac{1}{V_B}}{\frac{1}{V_{5'S}}} = \frac{V_{5'S}}{V_B}$$

or

$$RSE = \frac{\sigma_{5'S}^2}{\sigma_B^2} \times 100$$

which is the same as (6)

4.1.3 Relative Cost

It will be recalled (3.1) that when the data were taken the number of minutes needed for the interviewer to locate and count the large bolls on the

various plot sizes was recorded. In this study the cost of making boll counts on a sample section is expressed as the average number of minutes needed to make the boll counts on a 5'S portion of the plot, or, the unit of cost is the minute.

For example if it takes an average of 7.5 minutes for the bolls on 15'S units to be counted, the cost is

$$\frac{7.5}{(3)} = 2.5 \text{ minutes per } 5'S \text{ sub-unit}$$

Then, the cost of locating and making boll counts C_B on a unit of a given size B relative to the cost of locating and making boll counts $C_{5'S}$ on a 5'S sub-unit is

$$RC = \frac{C_B}{C_{5'S}} \quad . \tag{7}$$

4.1.4 Net Relative Efficiency

When considering a unit of a given size, B, and the "standard unit", 5'S, and using the ideas defined in 4.1.2 and 4.1.3, the net relative efficiency is

$$NRE = \frac{\text{RSE of B to 5'S}}{\text{RC of B to 5'S}}$$

$$= \frac{\frac{\sigma^2_{5'S}}{\sigma^2_B}}{\frac{C_B}{C_{5'S}}} \times 100 \tag{8}$$

The net relative efficiency, then, is directly proportional to the relative amount of information given by the boll counts on a plot of size B, and is inversely proportional to the relative cost of making the boll counts, as measured in minutes.

4.2 Analysis

The analysis was carried out in two phases. First, the mean squares were obtained by the analysis of variance (4.1.1.1) wherever possible, and the remaining three steps in the calculations were carried out (4.1.2, 4.1.3 and 4.1.4). Second, the alternative method (4.1.1.2) was used to get estimates of the mean squares, and the same remaining three steps in the calculations were again carried out.

4.2.1 First Phase

4.2.1.1 August

For the August data, which consisted of a complete enumeration of large bolls on a 25'D plot in each of eighty fields plus a large boll count on another 10'S plot in fifty out of eighty fields, two analyses of variance were carried out.

Using a table of random numbers one 10'S section out of the eight possible in the 25'D section was selected from each of the eighty fields.

Table III Analysis of Variance of August Data, Using 80 Fields

Source	d.f.	MS	E(MS)
Between fields	79	424.05	
Between 5'S in 10'S	80	42.62	σ_w^2

Then, using the data for the fifty fields in which both the 25'D and 10'S appeared, and again choosing one 10'S at random from the 25'D, the following was obtained,

Table IV Analysis of Variance of August Data, Using 50 Fields

Source	d.f.	MS	E(MS)
Between fields	49	1421.39	
Between 10'S in fields	50	132.44	σ_b^2

Combining information from both analyses (Tables III and IV) gives the following result, the mean square for "between 10'S in fields" being divided by 2 to put it on a 5'S basis.

Table V Combined Analysis of Variance of August Data

Source	d.f.	MS	E(MS)
Between 10'S in fields	50	66.22	$\sigma_w^2 + 2\sigma_b^2$
Between 5'S in 10'S	80	42.62	σ_w^2

By 4.1.1.1 $\sigma^2_{5'S} = \sigma^2_w + \sigma^2_b$. Therefore the "between 5'S in fields" mean square is

$$42.62 + \frac{66.22 - 42.62}{2} = 42.62 + 11.80 = 54.42$$

The number of degrees of freedom here are 53 obtained by using the method of Satterthwaite (1946).

The three succeeding steps may be easily followed by referring to Table VI. After the mean squares are placed on a 5'S basis, the relative statistical efficiency of 10'S to 5'S, by (6) of 4.1.2, is

$$\frac{54.42}{66.22} \times 100 = 82.18$$

The costs, 3.13 minutes/5'S for the 5'S section and 1.94 minutes/5'S for the 10'S unit were obtained from the original data gathered in the field. By (7) of 4.1.3 the cost of 10'S relative to 5'S is

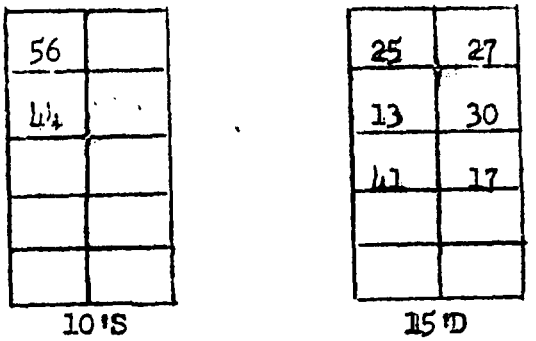
$$\frac{1.94}{3.13} = .620$$

Finally by (8) of 4.1.4 the net relative efficiency of 10'S to 5'S is

$$\frac{82.18}{.620} = 132.55$$

4.2.1.2 September

The September data consisted of fragmentary counts varying from 5'S to 25'D sections at two positions in each of forty fields plus a separate 10'S in each field. These data were treated similarly to the August data. However, for the larger plots, use could be made of the boll counts only in those fields in which both of the intended 25'D units contained enough information. For example, if interest were focused on getting the variance for "between 20'S in fields", a field of the type



where the blank spaces represent 5 foot lengths of single row on which the bolls were not counted, could not be used. Each such field represented the loss of a degree of freedom for estimating the variance. It is obvious that only a few of the fields out of the forty visited in September would happen by chance to contain enough information to allow that field to be used for estimating the variance between units of large size, say 20'D or 25'S, within fields.

However, an analysis of variance (4.1.1.1.) was run for all plot sizes, even if only three or four fields could be used. Results are shown on Table VI.

The steps leading to the net relative efficiency are computed in the same manner as described for August (4.2.1.1).

Table VI - A Summary of the Data Used in 4.2.1, to Arrive at the Net Relative Efficiencies of Various Plot Sizes, by Months.

Line Number	Type of s.u.	d.f.	Relative Size-B	Variance	MS on basis of 5'S. σ^2_B	Relative Statistical Efficiency	Cost 1/	Relative Cost	Net Relative Efficiency
	August								
(1)	5'S	53	1	54.42	54.42	100.00	3.13	1.000	100.00
(2)	10'S	50	2	132.44	66.22	82.18	1.94	.620	132.55
	Sept.								
(3)	5'S	23	1	47.95	47.95	100.00	2.34	1.000	100.00
(4)	10'S 2/	19	2		79.70	60.16	1.67	.714	84.25
(5)	10'S 2/	48	2	272.00	136.00	35.26	1.67	.714	49.38
(6)	15'S	12	3		70.20	68.30	1.25	.535	127.66
(7)	20'S	5	4	235.00	58.75	81.60	1.00	.427	191.10
(8)	25'S	3	5	818.30	163.66	29.30	1.07	.455	64.40
(9)	5'D	8	2	119.10	59.55	80.52	1.50	.641	125.62
(10)	10'D	4	4	59.00 3/	29.50	162.54	1.17	.498	326.39
(11)	15'D	4	6	130.00 3/	65.00	73.77	.78	.334	220.87
(12)	20'D 2/	3	8	15.30 3/	7.65	627.80	.71	.301	2085.71
(13)	20'D 2/	3	8	30.70 4/	7.68	624.35	.71	.301	2074.25
(14)	25'D	2	10	34.80 3/	17.40	275.57	1.17	.498	553.35

1/ minutes / 5'S sub-unit

2/ The two values for 10'S are from different analyses, as are the two values for 20'D.

3/ on 5'D basis

4/ on 10'D basis

4.2.2 Second Phase

4.2.2.1 August

Here the alternative method (4.1.1.2) was used to obtain the desired variances. An outline of the method of calculation is as follows.

To calculate, for example, a value for $\sigma^2_{25'S}$ for August, using the 25'D and 10'S units in each of fifty fields,

(1) One of the two 25'S units in each of the fifty 25'D units is chosen at random. Its boll count is x_1.

(2) Each of the fifty 10'S units, $x_{2'}$, is multiplied by $k = 5/2$. In each field $kx_{2'} = \hat{x}_2$

(3) The value of

$$\frac{\sum^{n}(x_1 - \hat{x}_2)^2}{2n}$$

is calculated, with $n = 50$.

(4) $(x_{2'})_1$ and $(x_{2'})_2$, each 10'S in size, are chosen at random from x_1, and at $k = 5/2$

$$\frac{\sum^{n}(x_{2'})_1 - (x_{2'})_2{}^2}{2n} = \sigma^2_{2'}$$

is calculated.

(5) The values of steps (3) and (4) are placed in equation (5) in 4.1.1.2, and σ^2_B is obtained by subtraction. This is the desired value, $\sigma^2_{25'S}$.

The steps leading to the value for net relative efficiency are carried out the same as in 4.2.1.1, and all results are shown in Table VII. Note that the degrees of freedom are now increased to fifty.

Table VII - A Summary of the Data Used in 4.2.2.1 to Arrive at the Net Relative Efficiencies of Various Plot Sizes for August.

Line Number	Type of s.u.	k	d.f.	Relative Size-B	Variance	MS on basis of $\frac{5}{2}$'S, σ_B^2	Relative Statistical Efficiency	Cost 1/	Relative Cost	Net Relative Efficiency
	August									
(1)	5'S 2/		53	1	54.42	54.42	100.00	3.13	1.00	100.00
(2)	10'S 2/	1	50	2	132.44	66.22	82.18	1.94	.620	132.55
(3)	15'S	$\frac{3}{2}$	50	3	157.50	52.50	103.66	1.58	.505	205.27
(4)	20'S	2	50	4	229.88	57.47	94.69	1.63	.518	182.80
(5)	25'S	$\frac{5}{2}$	50	5	321.82	64.36	84.56	1.73	.552	153.19
(6)	10'D	2	50	4	212.45	53.11	102.47	1.63	.518	197.81
(7)	15'D	3	50	6	225.64	37.61	144.70	1.46	.467	309.85
(8)	20'D	4	50	8	366.67	46.83	116.21	1.35	.429	270.88
(9)	25'D	5	50	10	978.50	97.85	55.62	1.24	.395	140.81

1/ minutes / 5'S sub-unit

2/ From Table VI

4.2.2.2 September

As described previously, the September data consisted of a 10'S unit and two incompletely counted 25'D units in each field. The 10'S unit and one of the two 25'D units, the one on which a sufficient number of the ten 5'S sub-units had boll count values, were selected in each field and the alternative method of estimating the mean squares was used. The method of computation is the same as for August, and the results are shown in Table VIII.

Table VIII - A Summary of the Data Used in 4.2.2.2 to Arrive at the Net Relative Efficiencies of Various plot Sizes for September.

Line Number	Type of s.u.	k	d.f.	Relative Size-B	Variance	MS on basis of 5'S σ_B^2	Relative Statistical Efficiency	Cost 1/	Relative Cost	Net Relative Efficiency
(1)	5'S 2/		23	1	47.95	47.95	100.00	2.34	1.000	100.00
(2)	10'S 2/	1	48	2	272.00	136.00	35.26	1.67	.714	49.38
(3)	15'S	$\frac{3}{2}$	23	3	193.55	64.52	74.32	1.25	.535	139.92
(4)	20'S	2	18	4	336.30	84.08	57.03	1.00	.427	133.56
(5)	25'S	$\frac{5}{2}$	9	5	285.81	57.16	83.89	1.07	.455	184.37
(6)	5'D 2/	1	8	2	119.10	59.55	80.50	1.50	.641	125.62
(7)	10'D	2	19	4	108.61	27.15	176.61	1.17	.498	354.63
(8)	15'D	3	13	6	356.95	59.49	80.60	.78	.334	241.32
(9)	20'D	4	8	8	600.33	75.04	63.90	.71	.301	212.29
(10)	25'D	5	4	10	655.93	65.59	73.11	1.17	.498	146.81

1/ minutes / 5'S sub-unit

2/ From Table VI.

Table IX. A Summary of the Results for the September Data

Type of s.u.	Analysis of Variance d.f.	NRE	Alternative Method d.f.	NRE
15'S	12	127.66	23	138.92
20'S	5	191.10	18	133.56
25'S	3	64.40	9	184.37
10'D	4	326.39	19	354.63
15'D	4	220.87	13	241.32
20'D	3	2085.71	8	212.29
25'D	2	553.35	4	146.81

V. Discussion and Interpretation of Results

Since the variable measured in August, large bolls, is not directly comparable to the variable measured in September, large unopen bolls, the results for each of the months will be discussed separately.

5.1 August

The use of the analysis of variance for obtaining the σ_B^2 values needed for computing the net relative efficiencies was limited to only two plot sizes, 5'S and 10'S. However, by using the alternative method, variances for all the plot sizes except 5'S, 10'S and 5'D were estimated with fifty degrees of freedom each, so the discussion will be confined to results from this method.

In considering the August data (large bolls) the best plot size appears to be between 10'D and 20'D, the most likely one of those studied being 15'D (Table VII). Although the variances were estimated with fifty degrees of freedom it must be remembered that the σ_B^2 estimates were biased downward, as explained in the development of the alternative method, 4.1.1.2.

5.2 September

September results were obtained by both methods so that here they may be compared. The results are summarized in Table IX.

In the results obtained by getting σ_B^2 from the analysis of variance, the values of the net relative efficiencies for plots of size 20'D and 25'D may be disregarded because of the small number of degrees of freedom with which their variances were estimated. Thus, the optimum plot size by both methods is in the neighborhood of 10'D. The value of NRE = 354.63 is quite reliable since it is based on 19 degrees of freedom. This plot size holds for counts of large unopen bolls only.

VI. Summary, Conclusions and Future Work

6.1 Summary

This study was carried out for the purpose of finding a plot size in a field of cotton which will give boll count values with a minimum of variability for a given expenditure of funds. This is a necessary step in an objective procedure designed to predict cotton yield.

The project was sponsored by the Bureau of Agricultural Economics of the U. S. D. A., the North Carolina Crop Reporting Service, and the Institute of Statistics at North Carolina State College. The stimulus for the project was the fact that a poor 1951 forecast cost the nation's cotton farmers an estimated

$125 million. Field data were collected during the 1953 growing season in Johnston and Harnett Counties, North Carolina. The data used consisted of boll counts on small sample plots in fields and the amounts of time needed to locate the plots and make the counts. The data were analyzed by a standard and an alternative statistical procedure with the cost of locating the plots and making the counts being measured by time.

6.2 Conclusions

The conclusions reached are

(A) When considering large bolls (August), the optimum plot size lies between 10'D and 20'D. The best plot size of those studied is 15'D.

(B) When considering large unopen bolls (September), the optimum plot size is 10'D.

6.3 Future Work

(A). It would be of interest to see the results of a study carried out in the same manner as this one, but in a "normal" year, in an area where cotton is of primary interest and with all necessary data being collected. This would serve as a check on the validity of the present study, especially on the use of the alternative method for getting the variances.

(B) An infinite model was postulated in the derivation of the alternative method. A more realistic approach, perhaps, would be the use of a finite model.

(C) When determining an optimum plot size, the entire sampling system, including locating the field and the plots, should be considered, since the optimum plot size may vary for different sampling systems.

VII. Bibliography

Agricultural subcommittee report. 1952. Crop estimating and reporting services of the Department of Agriculture. Report and recommendations of a special subcommittee of the Committee on Agriculture of the House of Representatives, 82nd Congress, second session. U. S. Government Printing Office, Washington, D. C.

Cochran, W. G., 1939. The use of the analysis of variance in enumeration by sampling. Jour. Am. Stat. Assn. 34: 492-510.

_____ 1953. Sample survey techniques. John Wiley and Sons, New York.

Hameed, A. 1953. Choice of plot size in the objective estimation of corn yield. Unpublished thesis. Library, Iowa State College, Ames, Iowa.

Horvitz, D. G. 1949. Sampling methods. Statistical Laboratory, Iowa State College, Ames, Iowa. (Ditto).

Johnson, F. A., 1943. A statistical study of sampling methods for tree nursery inventories. Jour. Forestry 41, 9: 674-679.

Kock, E. J. and Rigney, J. A. 1951. A method of estimating optimum plot size from experimental data. Agronomy Journal, 43, 1: 17-21.

Mahalanobis, D. C. 1945-1946. Report on the Bihar crop survey: rabi season 1943 - 1944. Sankyha 7, 1: 29-106.

Robinson, H. F., Rigney, J. A., Harvey, P. H. 1948. Investigations in plot techniques with peanuts. Tech. Bul. 86. N. C. Agri. Expt. Sta.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. Biom. Bull. 2:110-114.

Smith, H. F. 1938. An empirical law describing heterogeneity in the yields of agricultural crops. Jour. Agri. Sci., 28:1-23.

Sukhatme, P. V. 1947. The problem of plot size in large-scale yield surveys. Jour. Am. Stat. Assn. 42:297-310.

Yates, F. 1935. Some examples of biased sampling. Ann.Eugen. 6:202-213.