Pre-harvest sampling of soybeans for yield and quality

Earl E. Houseman
U.S.D.A.

Charles R. Weber
U.S.N.R.

Walter T. Federer
U.S.D.A.

Follow this and additional works at: http://lib.dr.iastate.edu/researchbulletin

Part of the Agriculture Commons, and the Statistics and Probability Commons

Recommended Citation
Available at: http://lib.dr.iastate.edu/researchbulletin/vol27/iss341/1

This Article is brought to you for free and open access by the Iowa Agricultural and Home Economics Experiment Station Publications at Iowa State University Digital Repository. It has been accepted for inclusion in Research Bulletin (Iowa Agriculture and Home Economics Experiment Station) by an authorized editor of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Pre-Harvest Sampling of Soybeans for Yield and Quality

By Earl E. Houseman, Charles R. Weber and Walter T. Federer

AGRICULTURAL EXPERIMENT STATION
IOWA STATE COLLEGE OF AGRICULTURE AND MECHANIC ARTS

STATISTICAL SECTION

BUREAU OF AGRICULTURAL ECONOMICS;
BUREAU OF PLANT INDUSTRY,
SOILS AND AGRICULTURAL ENGINEERING,
UNITED STATES DEPARTMENT OF AGRICULTURE

Cooperating

AMES, IOWA
CONTENTS

Summary ............................................................... 808
Introduction ......................................................... 809
Sampling procedure and data collected ......................... 810
Analysis of crop meter data ....................................... 814
Sampling errors and estimates from the sample .............. 815
Choice of subsampling unit ....................................... 820
Alternative sampling procedures ................................. 823
Forecasting aspects ................................................. 824
Literature cited ...................................................... 826
SUMMARY

The route-sampling method of estimating crop production has been extended to soybeans in a preliminary survey which is reported here. In 1941, just prior to harvest, 67 fields in eight east central Illinois counties were sampled for yield, percent protein, percent oil and iodine number of the oil.

Protein percent, oil percent and iodine number of the oil can be estimated satisfactorily, but estimating yield is more uncertain pending the accumulation of information on adjusting for harvesting losses and other factors which cause the sample average yield to be too large.

The yield of seed per acre differed with the method of planting (width of rows), indicating for the season studied that soybeans should have been planted in rows about 2 feet apart. The iodine number of the oil was lower for fields with wide rows than for drilled fields. This was attributed to difference in date of planting rather than method of planting.

It was concluded that two subsampling units should be taken per field and that the optimum size of subsampling unit is approximately 7 square feet.

Other investigations have shown that after the pods are fully distended there is little or no change in yield or chemical composition, indicating that production and quality can be estimated well in advance of harvest.
Pre-Harvest Sampling of Soybeans for Yield and Quality

BY EARL E. HOUSEMAN, CHARLES R. WEBER, AND WALTER T. FEDERER

The rapid increase in soybean production is due to versatile uses of the crop and the ease with which it fits into crop rotations in the Corn Belt. Soybean oil and oilmeal are used in food products, livestock feed and in industrial products such as paints, plastics and rubber substitutes. In anticipation of an increasing demand for pre-harvest information on soybean production and quality, a survey was conducted in 1941 just prior to harvest in order to test the applicability of route-sampling to soybeans. This survey covered eight counties in east central Illinois as indicated by the sampling routes in fig. 1.

Accurate estimates of production and quality would help the agronomist in ascertaining the reaction of varieties to particular geographical areas and climate and would be of value to the processor and farmer. At present, forecasts and estimates of yield and acreage are based on questionnaires distributed by mail, and no estimates of chemical composition are being made. It appears unlikely that farmers can accurately judge prior to harvest the yield of their soybeans. Adverse weather at the beginning of bean formation may cause the plant to adjust its bearing capacity by aborting seed or dropping part of its developing pods and beans.

Date of planting, environment and variety are among the factors affecting chemical composition. Planting date has been
found (7) to affect oil percentage and the iodine number of the oil. According to Carter and Hopper (1), oil formed under warm conditions tends to be more saturated or to have a lower iodine number than that formed under cooler environment. If yield and chemical composition of soybeans are to be estimated, the need for objective methods of sampling and estimation is clear.

**SAMPLING PROCEDURE AND DATA COLLECTED**

A practical method of making a random selection of fields has not yet been proposed and tested; however, even with a random selection of fields and a random sample within the fields, we still have the problem of adjusting the sample for harvesting losses. The route-method as described below provides a good workable way for getting an objective sample, but it does not give a random sample of all soybean fields and the sampling is restricted to the front portions of fields.

An attempt was made to select routes, fig. 1, that would give a good coverage of the counties sampled. Main highways were usually selected as they were easier to travel, especially in this case because of exceptionally heavy rains at the time the sample was taken.

Unfortunately the samplers carried part of the samples in a wet condition in their car for about 4 days, as provision had not been made for immediate drying of them. As a result a few beans started to mold, but all moldy beans were sorted out and not used in the chemical analyses. The exact effect of this on the chemical composition is not known.

In sampling, the crop meter, which is an instrument for recording the crop frontage along the route, serves two purposes: (i) It enables one to distribute the sample proportionally to soybean acreage, and (ii) it can be used to estimate soybean acreage. Hendricks (3), working with corn, alfalfa and wheat acreage in eight widely scattered counties, found that the areas of the fields along the roads tend to be proportional to their frontage. If this is also true for soybeans, the sample should have been distributed approximately in proportion to the soybean acreage because the number of fields sampled was kept proportional to the soybean frontage. In addition, the length of route in each county was made proportional to the land area of the county, which tends to stratify the sample by counties.

Upon entering a county the first field on the right side of the road was sampled except when the route continued directly from one county to the next, in which case the first field was determined from the crop meter. The car was stopped for
Fig. 1. Sampling routes of pre-harvest soybean survey, October, 1941.

Sampling each time the crop meter dial for the right hand side of the road showed that exactly 2 miles of soybean frontage had been passed. Actually, the crop meter was used to determine an objective point for entering a field as well as the field itself. This procedure of taking a sample at every 2 miles of frontage
was the device used for keeping the number of samples distributed in proportion to the soybean acreage. Checks made after the survey on the number of fields actually sampled in each county showed that the number was proportional to the soybean acreage.

In each field samples were taken at two locations which were determined as follows: The two samplers started in opposite directions along the edge of the field from a point in line with the windshield of the car. On a paper container for the pods, carried by each sampler, two random numbers between 0 and 100 had been written. Each sampler walked along the edge of the field a number of paces equal to his first random number and then proceeded into the field a number of paces equal to the second random number. The random number of paces taken into the field probably should have been between 10 and 110 in order to eliminate border effect.

To gain information on the best size and shape of subsampling unit, the beans at each location in the field were collected separately on six 1-foot × 1-row units situated as indicated in fig. 2. The reason for using an L-shaped unit was to provide a comparison of the variability between rows with the variability within rows, which would aid in deciding the best shape of subsampling unit.

From the end of the sampler's shoe on his last step into the field, an imaginary line perpendicular to the nearest row was visualized. The point 0 in fig. 2 corresponds to the point of intersection of this line and the nearest row. The pods were stripped from the stalks in each of the six 1-foot × 1-row units (lettered A, B, C, D, E and F) and put into small paper sacks. When the sampling at a location was completed the six sacks were placed in a larger container which was labeled as to field, variety and method of planting. Threshed weight (air-
dried) of the beans from each 1-foot $\times$ 1-row unit was converted to bushels per acre before making any statistical analyses.

Quality determinations were not made for each 1-foot $\times$ 1-row unit, because the quantity of beans was too small, and a chemical analysis of such a large number of samples would be rather costly. Thus, for chemical analyses the beans from units within a location were pooled.

Row widths in the fields sampled varied from 6 to 40 inches, and some fields were broadcast. In a few of the drilled fields it was impossible to determine the drill rows at one of the locations to be sampled. In that event, or in case the field was broadcast, a sampling unit 2 feet square was used. For purposes of analysis, the various row widths were classified into three groups: (i) narrow rows, (ii) medium rows and (iii) wide rows. The first group contained widths from 6 to 10 inches; the second, 20 to 28 inches; and the third, 36 to 40 inches. Of the 67 fields sampled, 12 had wide rows, 7 had medium width rows, 46 had narrow rows and two fields were broadcast. The two broadcast fields were classified in the narrow row-width group.

The frequencies of various varieties found in the fields sampled were: Illini 51, Dunfield 6, Manchu 6, Mandarin 1, Richland 1 and mixtures 2. Ten of the fields classified as Illini were predominantly Illini but contained various other varietal mixtures.

As indicated before, when the car was stopped, one observation was taken at random in each of two adjoining square areas which were each 100 $\times$ 100 paces in size. Regarding the sampling method as that of subsampling, the primary sampling unit is a segment or block of soybean acreage 200 paces long and 100 paces wide or deep along the right side of the road. The primary sampling units selected for the sample were spaced equally in terms of soybean frontage. Each primary sampling unit consists of two equal areas, and one secondary or subsampling unit is selected at random from each. The subsampling unit is a composite of six 1-foot $\times$ 1-row units. For convenience a primary sampling unit will be referred to as a field.

A distinction should be made between the entire soybean population and the population actually sampled, the latter being visualized as a strip 100 paces wide or deep along one side of the route. The problem can be viewed as one of sampling a strip of soybeans along a route and adjusting the sample for any difference between the population of soybeans sampled and the entire population, for harvesting and threshing losses, or for other factors.
ANALYSIS OF CROP METER DATA

If the expected value of the ratio of soybean frontage to the length of route is equal to the proportion of the total land area that is in soybeans, the frontage ratio expanded by land area should give an unbiased estimate of the total soybean acreage. This assumption did not prove to be correct. For the crops and counties studied, Hendricks (3) observed that the frontage ratio was not a direct measure of the proportion of land in a given crop because of a tendency for certain crops to be planted nearer highways than others. The proportion of land in a given crop is estimated as $kr$, where $r$ is the frontage ratio and $k$ is a constant which will be called the proportionality factor. In general the factor of proportionality cannot be assumed equal to unity and must be evaluated for each crop and area under inquiry before making acreage estimates, unless the same route is followed year after year and the estimates are made on a percentage change basis.

Expansions of the frontage ratio for each county by its land area are compared with the soybean acreages from the Illinois State Census in columns 2 and 3 of table 1. It is evident that the proportionality factor is greater than 1 since the unadjusted indications are consistently below the true acreages. From these data the best estimate of the proportionality factor is $(670,000/519,100) = 1.2907$ which was used to calculate the adjusted sample indications in column 4 of the table. It should be remembered that the choice of routes (i.e. whether or not

<table>
<thead>
<tr>
<th>County</th>
<th>Percent of total land in soybeans</th>
<th>Unadjusted sample indications</th>
<th>State census</th>
<th>Adjusted sample indications and estimated sampling standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Champaign</td>
<td>25.6</td>
<td>134000</td>
<td>163600</td>
<td>173000 ± 18900</td>
</tr>
<tr>
<td>Christian</td>
<td>24.4</td>
<td>91600</td>
<td>110800</td>
<td>118200 ± 16800</td>
</tr>
<tr>
<td>Ford</td>
<td>9.7</td>
<td>30700</td>
<td>30300</td>
<td>39600 ± 7400</td>
</tr>
<tr>
<td>Iroquois</td>
<td>9.5</td>
<td>41700</td>
<td>68000</td>
<td>53800 ± 5800</td>
</tr>
<tr>
<td>Kankakee</td>
<td>11.3</td>
<td>36000</td>
<td>49300</td>
<td>46300 ± 5500</td>
</tr>
<tr>
<td>Livingston</td>
<td>7.3</td>
<td>39500</td>
<td>48000</td>
<td>51000 ± 5500</td>
</tr>
<tr>
<td>Macon</td>
<td>23.1</td>
<td>62000</td>
<td>85200</td>
<td>80000 ± 11000</td>
</tr>
<tr>
<td>Vermilion</td>
<td>19.9</td>
<td>83600</td>
<td>114200</td>
<td>107900 ± 12600</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>16.1</strong></td>
<td><strong>519100</strong></td>
<td><strong>670000</strong></td>
<td><strong>670000</strong></td>
</tr>
</tbody>
</table>
they follow main highways and go through many towns or non-agricultural areas) affects the proportionality factor.

To obtain an estimate of the sampling error of the acreage estimate, crop meter readings were recorded every 10 miles in 6 of the counties. These readings provide estimates of the standard deviations of the soybean frontage per 10 miles, which can be converted to estimates of the standard error of the indicated acreages, assuming that the 10-mile segments were a random selection of all possible 10-mile segments. Only one of the adjusted acreages differs significantly from the state census. The same proportionality factor appears appropriate for all of the counties.

The coefficients of variation of soybean frontage per 10 miles for the 6 counties showed no tendency to be correlated with the varying proportion of total land in soybeans which ranged from about 7 to 26 percent among counties; i.e., the standard deviation of frontage per 10 miles was proportional to the average frontage per 10 miles. The average coefficient of variation was 45 percent.

In the 8 counties sampled the length of route was 1070 miles. Hence, the sampling standard error of the indicated total soybean acreage is estimated at 4.4 percent, (45/√107), again assuming no errors in expansion and regarding the 10-mile segments of the route as a random selection of such segments.

**SAMPLING ERRORS AND ESTIMATES FROM THE SAMPLE**

For purposes of analysis the sample is regarded as a stratified random sample of fields, the stratification being by counties and the fields subsampled. Hence the variance among fields within counties is taken as the sampling error. Actually, the fields sampled are at equal intervals of soybean frontage, which is expected to give a more accurate sample of all fields along the route than a random selection. It is therefore believed that the sample mean square among fields within counties is a positively biased estimate of the sampling error; however, a counteracting factor is an error of unknown magnitude attributable to differences in routes that might be chosen. In the absence of further knowledge, the mean square among fields within counties appears to be the most satisfactory estimate of the sampling error.

---

The standard errors of the acreage indications were estimated as

\[
\frac{L \cdot k \cdot s}{10\sqrt{n}},
\]

where \(L\) = land area, \(k\) = proportionality factor, \(s\) = standard deviation of soybean frontage expressed in miles per 10 road miles and \(n\) = number of 10-mile segments.
Analyses of variance of yield in bushels per acre, of protein percentage, of oil percentage and of the iodine number of the oil are presented in table 2. The mean squares are on an individual subsampling unit basis. Thus, the estimated sampling standard errors per mean of two subsampling units per field are calculated by dividing the mean squares between fields within counties by 2 and extracting the square root. This gives: yield, 7.28; protein percentage, 2.44; oil percentage, 1.00; and iodine number, 2.76. Division of these sampling standard errors by the square root of 67, the number of fields sampled, gives the estimated sampling standard errors of the sample averages.

**TABLE 2. ANALYSES OF VARIANCE OF YIELD, PERCENT PROTEIN, PERCENT OIL AND IODINE NUMBER OF THE OIL.**

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Yield</th>
<th>Percent protein</th>
<th>Percent oil</th>
<th>Iodine number</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d.f.</td>
<td>m.s.</td>
<td>d.f.</td>
<td>m.s.</td>
</tr>
<tr>
<td>Counties</td>
<td>7</td>
<td>190</td>
<td>7</td>
<td>21.8</td>
</tr>
<tr>
<td>Fields within counties</td>
<td>59</td>
<td>106</td>
<td>59</td>
<td>11.9</td>
</tr>
<tr>
<td>Locations within fields</td>
<td>67</td>
<td>76</td>
<td>65*</td>
<td>2.7</td>
</tr>
</tbody>
</table>

*Parts of two fields are missing.

Because of the desirability of having a crew of two men per car, a driver and an operator of the crop meter, it is assumed that not less than two observations should be taken per field. Previous investigations on sampling wheat (5, 6) and sampling corn (2) have shown that the amount of sampling per field should be kept at a minimum in order to sample as many fields as possible, as the variability among fields was considerably larger than the variability within fields.

In the study on soybeans, the mean square of yield within fields was about three-fourths as large as that among fields within counties (table 2). This suggests that it might be profitable to enumerate more than two subsampling units per field, as less time is required to double the amount of sampling per field than to double the number of fields sampled. The results of this study indicated practically no choice between two subsamples per field and four subsamples per field with one-half as many fields. Therefore it is concluded that two subsamples are the best number to take per field when estimating the yield of soybeans. In substantiation of this conclusion the results for quality (iodine number, protein percentage and oil percentage) indicated that the variability within fields was several times smaller than the variability among fields. From
the time records that were kept it was clear that in sampling for quality, additional time would be more profitably spent by stopping more frequently along the route than by increasing the amount of sampling per field. Hence the results on soybean sampling indicating that the number of samples per field should be kept at the minimum, 2, are in agreement with the studies on wheat (5, 6) and corn (2).

The sample averages with their standard errors for the four soybean characters studied are: yield, 29.5 ± 0.9 bushels per acre; protein, 41.6 ± 0.3 percent; oil, 20.9 ± 0.1 percent; and iodine number, 129.6 ± 0.3. The yield estimate for the sample average, 29.5 bushels per acre, is 27 percent higher than the estimate of 23.3 bushels per acre obtained from county statistics published by the Bureau of Agricultural Economics. There were no apparent reasons to expect any bias in the estimates for quality of soybeans; hence the discussion will be confined to a probable explanation of the difference in the two estimates for yield of soybeans.

Four sources for the difference in the two estimates were considered; these are as follows:

1. Harvesting and threshing losses.
2. Difference in the concept of soybean acreage.
3. Difference in the acreage and yield of soybeans along main highways and the total soybean acreage.
4. Tendency to sample an area larger (or smaller) than specified.

The harvesting and threshing losses probably account for one-half or more of the difference in the two estimates. Owing to the heavy rains, the soybean crop in the area sampled was not harvested until many fields had been mature for a considerable time. Thus adverse weather conditions and shattering due to overmaturity were conducive to considerable harvesting losses. Hurst and Humphries (4) presented data which show that on the average harvesting and threshing losses appear to be a constant percentage of the yield. Hence it appears that eventually a satisfactory adjustment factor can be worked out for harvesting and threshing losses.

The second factor causing a difference in the two estimates is the concept of an acre of soybeans. The farmer quite likely reports an acreage which includes nonproductive areas along the edge of the field or (and) small areas of noncropable land within the field. Such areas are excluded by the sampling method; however, in case an observation point falls in a portion of a field where the crop has been destroyed, the observation is counted as zero in estimating the yield but is not counted when estimating quality.

There is little or no knowledge concerning the third and
fourth factors given above. However, from the records kept for the individual samplers, no difference between yields corresponding to samplers could be detected. This indicated that both samplers included very nearly the same area in the sample. Either both samplers included a larger area than specified or else they included the correct area specified.

Important differences in soybean yield for the various row widths were observed as indicated by the analysis of variance in table 3. The three row-width groups are as described before, the broadcast fields being included in the narrow row-width group. All fields sampled in one of the counties (Christian) were in the narrow row-width group. This county had a comparatively low sample average yield, 24.3 bushels, which did not appear attributable to row width, and was omitted from the comparisons in order to avoid a possible confounding of effects associated with county and of effects associated with row width. Average yields in bushels per acre by row width for the 7 counties analyzed were: narrow, 31.4 (38); medium, 34.4 (7); and wide, 24.7 (12); where the numbers in parenthesis refer to the number of fields sampled. As seen from table 3 the county effects were small or non-existent so there is little or no confounding of county and row-width effects. Hence the yields for the different row widths are directly comparable.

### Table 3. Analysis of Variance of Yields Comparing Row Widths.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Mean square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row widths</td>
<td>2</td>
<td>531.3</td>
<td>6.11*</td>
</tr>
<tr>
<td>Counties within row widths</td>
<td>15</td>
<td>97.0</td>
<td></td>
</tr>
<tr>
<td>Fields within county within row widths</td>
<td>39</td>
<td>86.9</td>
<td></td>
</tr>
</tbody>
</table>

*F > .01 level of probability.

Iodine number of the oil also varies with row width as seen from table 4. Iroquois County is the only one in which the iodine number is higher for the wide rows than for the narrow rows, and in that ease the single observation for the narrow rows happened to be smaller than the average for any other county in that group. There is considerable belief that the narrow rows were planted later than the wider rows in order to control weeds. In addition, there is good supporting evidence that the iodine number of oil is increased as the date of planting is delayed (7). Therefore a plausible explanation for the broad-
cast and drilled fields having the higher iodine number is date of planting.

The iodine number for the two varieties, Illini and Manchu, was about the same whereas the iodine number for Dunfield

### TABLE 4. ESTIMATES OF IODINE NUMBER OF THE OIL BY ROW WIDTHS AND ANALYSIS OF VARIANCE FOR ILLINI.

<table>
<thead>
<tr>
<th>Variety and county</th>
<th>Narrow*</th>
<th>Medium*</th>
<th>Wide*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illini</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macon</td>
<td>(8) 130.8</td>
<td></td>
<td>(1) 129.4</td>
</tr>
<tr>
<td>Christian</td>
<td>(9) 129.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Livingston</td>
<td>(2) 130.5</td>
<td>(1) 128.9</td>
<td>(2) 125.2</td>
</tr>
<tr>
<td>Iroquois</td>
<td>(1) 128.6</td>
<td>(1) 130.4</td>
<td>(1) 129.7</td>
</tr>
<tr>
<td>Kankakee</td>
<td>(2) 133.1</td>
<td>(1) 130.0</td>
<td>(1) 129.7</td>
</tr>
<tr>
<td>Ford</td>
<td>(1) 133.0</td>
<td>(1) 129.8</td>
<td>(2) 129.1</td>
</tr>
<tr>
<td>Vermilion</td>
<td>(6) 131.3</td>
<td>(1) 129.6</td>
<td></td>
</tr>
<tr>
<td>Champaign</td>
<td>(9) 130.9</td>
<td>(1) 130.0</td>
<td></td>
</tr>
</tbody>
</table>

Average

| Manchu            | (3) 130.2 | (1) 131.6 | (2) 129.1 |
| Dunfield           | (4) 126.0 |         | (2) 122.5 |

### ANALYSIS OF VARIANCE (ILLINI ONLY)

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Mean square</th>
<th>F-ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row widths</td>
<td>2</td>
<td>70.1</td>
<td>4.64†</td>
</tr>
<tr>
<td>Counties within row width</td>
<td>15</td>
<td>19.7</td>
<td></td>
</tr>
<tr>
<td>Fields within county within row width</td>
<td>33</td>
<td>15.1</td>
<td></td>
</tr>
</tbody>
</table>

*Numbers in parentheses refer to number of fields sampled.
†F > .05 level of probability.

was lower, which agrees with findings reported by Cartter and Hopper (1). As was the case with yields, the mean square for counties within row widths was of the same order of magnitude as the mean square for fields within counties within row widths, table 4, indicating no confounding of county and row width effects.

Analyses of variance of protein and oil content are given in table 5. As protein and oil percentage did not vary with row width, a classification by row width is not shown in the table. The three varieties, however, differed as to protein percentage and oil percentage, and when ranked according to either were entirely consistent with results presented by Cartter and Hopper (1).
TABLE 5. ESTIMATES OF PERCENT PROTEIN AND PERCENT OIL BY SOYBEAN VARIETIES AND ANALYSES OF VARIANCE

<table>
<thead>
<tr>
<th>Variety</th>
<th>Number of fields</th>
<th>Percent protein</th>
<th>Percent oil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illini</td>
<td>51</td>
<td>41.4</td>
<td>20.9</td>
</tr>
<tr>
<td>Manchu</td>
<td>6</td>
<td>44.1</td>
<td>20.2</td>
</tr>
<tr>
<td>Dunfield</td>
<td>6</td>
<td>39.7</td>
<td>22.2</td>
</tr>
</tbody>
</table>

ANALYSES OF VARIANCE.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>d.f.</th>
<th>Protein mean square</th>
<th>Oil mean square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varieties</td>
<td>2</td>
<td>58.5*</td>
<td>12.65*</td>
</tr>
<tr>
<td>Counties within varieties</td>
<td>12</td>
<td>14.6</td>
<td>2.08</td>
</tr>
<tr>
<td>Fields within county within variety</td>
<td>48</td>
<td>10.5</td>
<td>1.65</td>
</tr>
</tbody>
</table>

*F > .05 level of probability.

CHOICE OF SUBSAMPLING UNIT

In calculating an estimate of the optimum size of subsampling unit a fixed length of route is assumed. The problem is then stated as one of determining the optimum size when the time saved or expended by changing the size of the subsampling unit is used to stop more or less frequently along the route. The data collected and various surrounding circumstances are such that only rough calculations are worth making.

The variance of the sample average is proportional to

\[ \frac{(t_1 + kt_2)(A + B)}{k} \]

The quantity \( t_2 \) is the average time spent per unit area (1 square foot) in stripping the pods from the plant, whereas \( t_1 \) represents the average time required to sample a field apart from the time used in gathering the pods. Since the length of route is assumed constant, \( t_1 \) is the average time used in getting to and from the subsampling units, not including time spent in traveling. Thus, the first factor, \( t_1 + kt_2 \), represents an average time to sample a field when subsampling units of \( k \) square feet are used.

The second factor, \( A + \frac{B}{k} \), is the variance between fields and is on a basis equivalent to the mean of \( k \) 1-foot-square units. \( A \) is the variance attributable to fields plus the variance between
locations within fields; i.e., the variation other than the component of sampling error arising at the points where the subsampling units are taken, assuming two subsampling units per field. $B$ represents the variance per square foot within a subsampling unit of undefined size. Actually, $A$ and $B$ are functions of $k$, the size of subsampling unit; however, it appears sufficient to assume $A$ and $B$ independent of $k$ if the optimum size of subsampling unit calculated under that assumption is about the same size as the average size of subsampling unit actually used in the survey. The value of $k$ which will minimize the sampling variance is $\sqrt{\frac{t_1 B}{t_2 A}}$.

Analyses of variance of yields in bushels per acre by row widths are presented in table 6. These analyses are on the basis of a mean of six 1-foot $\times$ 1-row units, so the mean squares between fields within counties are estimates of a quantity of the form of $A + \frac{B'}{6}$. The value of $B'$, which is the variance between 1-foot $\times$ 1-row units within the subsampling units, varies with row width. As the mean squares within subsampling units are not comparable, they were converted to a square foot basis and are shown in the bottom row of table 6. Weighting the individual estimates of $A$ and $B$ from each row width group by the degrees of freedom, we have 34.6 and 291 as estimates of $A$ and $B$ respectively.

Unfortunately the data collected on the time to perform various operations do not permit an accurate estimate of $t_2$. The total time spent at each subsampling unit was recorded and from these records an average time per square foot was calculated, but this overestimates $t_2$ because of time required in keeping the 1-foot $\times$ 1-row units separate. The desired value of $t_2$ is simply the average time required to strip the pods from plants in an additional square foot and put them into the same container. The present data indicate that this time is about 1 minute, but that should be checked at a later date. It is expected that $t_2$ will vary somewhat from year to year depending upon the condition of the plants (e.g., the number of leaves on them) at the time of the survey. At present the best estimate of $\frac{t_1}{t_2}$ is 6, which gives $\sqrt{(6)(291/34.6)} = 7.1$ square feet (approximately $1/6000$ of an acre) as the optimum size of subsampling unit. This is a little smaller than 7.4, the average size for the 60 fields used in the analyses presented in table 6. In the wheat surveys the units used were about $1/5000$ of an acre or 8.7 square feet (6) and $1/10,000$ of an acre (5).
### TABLE 6. ANALYSES OF VARIANCE OF YIELD BY ROW WIDTHS.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Narrow rows</th>
<th>Medium rows</th>
<th>Wide rows</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>d.f.</td>
<td>m.s.</td>
<td>Quantity estimated</td>
</tr>
<tr>
<td>Fields within counties</td>
<td>35</td>
<td>93.5</td>
<td>$A + \frac{B}{6}$</td>
</tr>
<tr>
<td>Subsampling units within fields</td>
<td>42</td>
<td>99.5</td>
<td>$\frac{B}{6}$</td>
</tr>
<tr>
<td>Within subsampling units</td>
<td>420</td>
<td>75.3</td>
<td>$\frac{B}{6}$</td>
</tr>
<tr>
<td>Mean square within rows (A, B, C)†</td>
<td>168</td>
<td>65.9</td>
<td>28</td>
</tr>
<tr>
<td>Mean square across rows (D, E, F)†</td>
<td>168</td>
<td>83.5</td>
<td>24*</td>
</tr>
<tr>
<td>Average width of row in feet</td>
<td>0.597</td>
<td>1.84</td>
<td>3.26</td>
</tr>
<tr>
<td>Mean squares per square foot within subsampling units</td>
<td>270</td>
<td>350</td>
<td>338</td>
</tr>
</tbody>
</table>

*Part of one subsampling unit was accidentally destroyed.
†The mean squares within and across rows were divided by 6 to make them comparable with the mean squares in the first part of the table.
A unit of about 7 square feet is practical and gives a reasonable quantity of material to handle. It is recommended that subsampling units of approximately the same size be used in each field. For the broadcast and drilled fields a U-shaped frame works better than a measuring stick. Most of the drill rows are 7 inches wide in which case an area 4 rows × 3 feet is 7 square feet. A U-shaped frame that could be used for all drilled and broadcast fields and the area covered in each case are illustrated in fig. 3. It is necessary to keep a record of the row width of each field, as a different conversion factor must be used for each width.

In the fields with the wider rows it is suggested that the sampling be confined to a single row, as the results in table 6 indicate there is little difference in variability between and within rows. For the wide rows the sampling frame is not used, and one can either sample a fixed length of row or vary the length of row sampled so the same conversion factor can be applied regardless of row width. The latter is easily accomplished by having lengths, corresponding to each row width, marked on the measuring stick. For a subsampling unit of 7 square feet the length of row would vary from 50.4 inches with a 20-inch row width to 25.2 inches when the width of rows is 40 inches.

**ALTERNATIVE SAMPLING PROCEDURES**

Yield per acre can be written as the product of three factors, \( p \), \( r \) and \( w \), where \( p \) is the average number of plants per acre, \( r \) is the average number of pods per plant and \( w \) is the average weight of beans per pod expressed in bushels. Each of these factors is subject to sampling error which contributes to the total sampling error of the yield estimate. When plans for the survey were being made, the question was raised as to the pos-
sibility of getting a more efficient sampling design by sampling for some factors at heavier rates than others.

It was immediately apparent that time should not be spent in making pod counts as the pods can be stripped from the plants faster than an accurate count of them can be made. Therefore, the factors $r$ and $w$ are combined and yield is considered as the product of $p$ and $w'$, where $w'$ is the average weight of beans per plant. If the relative sampling error (coefficient of variation) for individual plant yields should have been small in comparison with the relative sampling error for number of plants, a good sampling plan would be to count the number of plants on an area perhaps several times larger than the unit of area from which the pods are taken, especially if plant counts can be made faster than pods can be stripped. In sampling for corn yield a similar plan, in which ears are counted instead of plants, has been found practicable (2). During the soybean survey, however, only a limited number of plant counts were made, and it is not possible to investigate what might have been gained by such a scheme. The writers feel, in view of the large variability in plant yields, that modifying the sampling to include plant counts on a larger area and taking the pods from only a few plants to estimate the yield per plant is likely to be unprofitable, especially for drilled and broadcast fields.

**FORECASTING ASPECTS**

In view of the fact that these preliminary estimates of soybean production and quality have given encouraging results in that it was possible to sample soybeans objectively, the question might be asked, "How far in advance of maturity can reliable forecasts of production and quality be made by means of objective field counts?" This question can not be answered adequately as yet, but previous investigations, (1) and (7), have given indications of the answer. Data have provided evidence that fairly accurate forecasts of yield and quality might be made as early as the latter part of August. This stage of maturity of the soybean plant would be before any leaves have dropped due to normal maturation. Investigations (1) have shown that where physiological processes in the plant had reached a point when the pods were fully distended (about 4 weeks before harvest), there was no appreciable change in yield or in chemical composition between that date and harvest. On the basis of the above investigations, soybean production and quality could not only be forecast but estimated approximately 4 weeks in advance of harvest. Possibly yield, but not quality, could be forecast 5 to 6 weeks before harvest.
Sampling just after full distention of the pods is more difficult than sampling immediately before harvest. At full distention the leaves are still on the plants and it is necessary to pick the pods instead of stripping them from the stalks. In view of the greater time required to take the pods from the plants, the previous idea of making plant counts on a relatively large area and picking pods from a small number of plants might become feasible. An early sampling is also complicated by the fact that some fields will be cut for hay and might differ from fields to be harvested for grain.

Several possibilities of forecasting yield 5 or 6 weeks before harvest might be advanced. One is suggested when yield is considered as the product of the average number of seeds per acre, which can be obtained by sampling, and the average seed size at harvest over a period of years. The error of such a forecast would be attributable to the variation in average seed size from year to year and error in the estimate of the number of seeds per acre. The coefficient of variation among years for seed size (grams per 100 seeds), as calculated from table 4 in reference (1), is 6 percent. This result is based on 5 years’ data for the same varieties.

Favorable weather conditions during early stages of seed development might cause a larger than average number of seeds to set per plant, and unfavorable weather following seed setting would retard seed development. Thus the smallest average seed sizes are to be found in years when weather conditions favor the setting of a large number of seeds and the subsequent weather is adverse. On the other hand, adverse weather during seed setting which is followed by favorable developmental conditions for the remainder of the season leads to seeds of larger than average size. From a study of the relationship between final seed size and weather, one might be able to forecast the final outcome of seed size and obtain better results than using the same average seed size each year.
LITERATURE CITED


