Soybean Production in the Midsouth

Edited by

LARRY G. HEATHERLY
U.S. Department of Agriculture
Agricultural Research Service
Stoneville, Mississippi

HARRY F. HODGES
Mississippi State University
Department of Plant and Soil Sciences
Mississippi State, Mississippi

EDITORIAL COMMITTEE: RICHARD WESLEY AND ALAN BLAINE

SPONSORED BY THE MISSISSIPPI SOYBEAN PROMOTION BOARD

CRC Press
Boca Raton London New York Washington, D.C.

Acquiring Editor:	John Sulzycki
Project Editor:	Carol Whitehead
Marketing Manager:	Becky McEldowney
Cover design:	Dawn Boyd
PrePress:	Greg Cuciak
Manufacturing:	Carol Slatter

Library of Congress Cataloging-in-Publication Data

Soybean production in the midsouth / edited by Larry G. Heatherly and
Harry F. Hodges.
p. cm.
Includes bibliographical references and index.
ISBN 0-8493-2301-0 (alk. paper)
1. Soybean–Southern States. I. Heatherly, Larry G. II. Hodges, H. F.
SB205.S7S39 1998
633.3'4'0975–dc21 98-24743
CIP

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Direct all inquiries to CRC Press LLC, 2000 Corporate Blvd., N.W., Boca Raton, Florida 33431.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are only used for identification and explanation, without intent to infringe.

International Standard Book Number 0-8493-2301-0
Library of Congress Card Number 98-24743
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

chapter seventeen

Sampling tips and analytical techniques for soybean production

J. L. Willers, G.W. Hergert, and P. D. Gerard

Contents

Overview

Sampling is an extensively discussed topic in the culture of many crops including soybean. Thus, most sampling methods applicable to soybean production are already well described. Therefore, the primary focus of this chapter is to provide concepts that help implement a sampling plan and analyze and interpret its sample data. We believe teaching

0-8493-2301-0/99/$0.00+$.50

how to interpret sample data is the chief shortcoming of the sampling literature available to soybean producers. Our goal is to help the reader (i.e., a consultant, farm manager, or producer) better understand the sampling process and give guidance (through a few "hands-on" examples) on what to do with sample data acquired from soybean fields. To aid in this process, several sections provide descriptions of pencil and paper methods that can be applied to sample information. Applying these simple methods will build confidence and proficiency in sampling and interpreting skills. These skills will also help one to better read and understand a greater portion of the sampling literature available. The use of these analysis techniques may result in more insight (by identifying patterns) that could lead to the development of improved production methods. Several sets of data will be analyzed to illustrate key concepts about sampling. Some discussion is provided for a few of the more recent sampling concepts and techniques, in particular, those pertaining to site-specific management (SSM), otherwise known as precision agriculture.

The emphasis on analysis and the description of several hands-on methods is what makes this chapter unique. We have not, however, been able to avoid repeating elementary statistical concepts and some rudiments of sampling theory. We also introduce a modification of the line-intercept sampling (LIS) method, a procedure currently underutilized in row crop agriculture. The use of a graph, called an interaction (or effects) plot, is also described.

Sampling is an activity of obtaining information. The next major breakthrough in agricultural production will likely result from better applications of information. Sampling will be one of several processes at work to provide information that leads to better productivity. In summary, our hope is that this chapter will help the reader be more aware of the need to sample, become more skilled in the analysis of information, and be better informed to select and properly apply soybean sampling techniques.

Introduction

Sampling correctly is a difficult task to learn and one that can be time-consuming and monotonous. Another hindrance is the difficulty of mastering methods of analysis necessary for interpreting sample data. Nevertheless, the reader is invited to travel on a journey into the world of sampling and expected to have a learning experience. You will have to do some work, but once you have acquainted yourself with some key concepts, skill in the use of a valuable tool will have been acquired. Eventually, the time required to learn new sampling concepts will begin to decrease dramatically and improvements in the skills of getting better information for making production decisions will follow. It is hoped the practice of sampling will become more of a faithful companion rather than a distant land better heard about than traveled.

Proper applications of sampling techniques can help increase profits and improve management of the crop. On the other hand, poor sampling techniques that produce "bad" data can lead to erroneous conclusions and be expensive wastes of time and resources. Sampling correctly is a skill that must be mastered. Without learning how to sample, one may never discover improvements that produce better crops. This chapter provides information to identify the important literature on traditional sampling techniques useful for soybean production. These excellent sources are quite complete. To merely repeat their information here would be redundant. The chapter bibliography can be used to help locate these materials from the county extension office, local or university library, or bookstore by special order.

Basic concepts and principles of sampling

In this and following sections, words that have unique meaning appear in bold font. The concept or idea of that word, if it is not immediately described, should become clearer as the discussion proceeds. However, it is hoped that by calling them to your attention a mental model of the concepts and the relationships between them will begin to form in your mind.

Sampling is a specialized discipline of the field of **statistics**. Statistics has as its principal purpose the goal of summarizing collections of data into smaller sets of data (or information) so that relationships (as similarities or differences) between (or among) **populations** can be identified. Thus, the word *statistic*(s) can refer (1) to an academic field of study or (2) to a collection of one or more numbers (i.e., the smaller sets of data) used to summarize (or describe) the characteristics of a population.

Sampling is the application of a prescribed method to a population of individuals (insects, soybean plants, fields, farms, people, etc.) for the purpose of obtaining information (data) to describe that population. A particular method of sampling is called a **sampling design**. The basic task of sampling is to collect, without prejudice, a small sample of individuals from a much larger population of interest and measure from them one or more characteristics of interest. These results are then generalized to the larger population. Often, a mathematical equation, called a **probability density function** or pdf for short, can be used to describe different population traits that are sampled. Estimates are made from the sample to determine **parameters** that "fit" the pdf function to that particular population and not to some other. It is important to keep in mind the distinction between a **population parameter** and an **estimate** (i.e., statistic) of that parameter. The parameter is the true value of a **variable** (i.e., characteristic) when all individuals of a population are measured; that is, it is the value that actually exists if it could be measured. Since it is impossible (in many cases) to measure all individuals of a population, one must sample the population and estimate the value of a parameter from the sample. Later, examples are provided that illustrate fitting the **normal pdf** to different populations of soybean yields and plant heights.

The normal pdf is a common distribution used in many sampling applications. It is not, however, the only one that can be used; for example, the **lognormal** pdf is another common distribution. The binomial distribution is often used in sampling designs where the sample response is a "yes" or a "no" (i.e., does the selected plant in the sample have aphids or not?). More than a dozen other distributions can be applied in sampling designs.

Therefore, sampling involves the use of many concepts, including (1) how to select **sample units**, (2) determining the size (**sample unit size**) and (3) the number of sample units (or **sample number size**), (4) the specification of how often samples will be collected, (5) who will collect the samples, (6) the nature of the data collected, (7) how the data will be analyzed, and (8) to whom the data and analysis will be presented. A sampling design, or plan, describes how these traits are related to each other.

A sampling design should accomplish clear and well-defined objectives (Buntin, 1993). These objectives should be established before any data are collected. It is a good practice to enlist the aid of a statistician (or other expert) when developing a novel sampling plan to provide advice and guidance in its development. On the other hand, if you are the user of an already established sampling plan, you should take steps to ensure that the plan is correctly implemented. In short, learn and understand the assumptions of the sampling plan you choose to employ. The authors of that sampling plan should have provided this information in their description of that plan.

Different sampling plans are best suited to specific tasks; therefore, it is necessary to use a sampling plan that matches a specific need. This judgment can only be made when you know the strengths and limitations of alternative sampling plans. The ***assumptions*** that lie behind any particular sampling plan should be clear to you before you trust its results.

To help facilitate an understanding of the sampling literature, we next present concepts, definitions, and a few commonly used algebraic expressions. We will use several

Table 17.1 Selected Yields (bu/acre) for Maturity Group V Soybeans in the 1994 Mississippi Soybean Variety Trials

		Location				
Variety	Brand	Clarksdale	Rolling Fork	Hernado	Verona	Average ± *s*
A5560	Asgrow	32.9	50.8	77.7	48.6	52.50 ± 18.59
A5843	Asgrow	43.4	58.5	77.7	44.0	55.90 ± 16.12
A5885	Asgrow	40.9	58.2	75.8	53.5	57.10 ± 16.80
A5979	Asgrow	54.0	56.6	78.2	45.2	58.50 ± 14.01
Average ± *s*		42.80 ± 8.71	56.03 ± 3.58	77.35 ± 1.06	47.83 ± 4.26	
DPL 105	Deltapine	46.2	51.1	78.6	43.3	54.80 ± 16.19
DPL 415	Deltapine	48.4	54.6	73.5	48.1	56.15 ± 11.95
DP 3589	Deltapine	42.1	62.6	69.5	58.4	58.15 ± 11.64
Average ± *s*		45.57 ± 3.20	56.10 ± 3.89	73.87 ± 4.56	49.93 ± 7.71	
H5164	Hartz	44.6	49.3	75.5	46.3	53.92 ± 14.51
H5088	Hartz	45.7	55.4	72.8	51.8	56.42 ± 11.63
H5218	Hartz	46.7	49.1	79.4	47.5	55.67 ± 15.85
H5350	Hartz	44.0	48.9	77.1	37.8	51.95 ± 17.37
H5454	Hartz	49.3	57.0	77.9	51.1	58.82 ± 13.13
H5545	Hartz	54.2	52.3	72.6	43.1	55.55 ± 12.36
H5566	Hartz	43.6	48.2	74.9	42.0	52.17 ± 15.38
H5810	Hartz	44.2	42.0	67.4	42.9	49.12 ± 12.22
Average ± *s*		46.54 ± 3.61	50.27 ± 4.67	74.70 ± 3.78	45.31 ± 4.77	
9501	Pioneer	36.8	53.1	66.6	37.5	48.50 ± 14.22
9551	Pioneer	37.1	49.4	68.1	38.9	48.37 ± 14.22
9584	Pioneer	44.9	53.8	81.0	50.9	57.65 ± 16.00
9592	Pioneer	52.9	53.1	84.7	51.1	60.45 ± 16.19
9593	Pioneer	50.5	52.9	77.0	51.7	58.02 ± 12.69
Average ± *s*		44.44 ± 7.43	52.46 ± 1.74	75.48 ± 7.92	46.02 ± 7.16	
Overall average ± *s*		45.12 ± 5.62	52.84 ± 4.58	75.30 ± 4.71	46.68 ± 5.58	

sample data sets to illustrate the application of these principles. The comment of Cochran (1956) should be kept before you at all times. He stated, "...any sampling plan contains two parts: a rule for drawing the sample and a rule for making the estimates from the results from the sample." The preceding quote is a good summary of the concepts presented so far.

Basic definitions and tools

The first concept to focus on is what is meant by the word **population**. From a sampling perspective, a population is a collection of individuals (or items), such as a field of soybean plants. These individuals will have one or more characteristics, labels, or other criteria that exist and are shared by all individuals. The definition of what is meant by the word *population* can change as different sampling objectives are established. In some instances, the same individual can belong to more than one population depending upon the goals of a sampling activity.

For example, there are several populations that can be constructed from the data in Table 17.1. We could consider the population of each variety by itself (20 populations having four individuals or locations each), or the populations of varieties within brands (4 populations with 16, 12, 32, or 20 individuals each). Other populations that could be defined are the soybean varieties at any of the four locations (4 populations with 20 individuals each). This exercise should make clear that the population of interest should be clearly defined as determined by the question(s) or objectives of interest.

Table 17.1A Frequency Distribution of the Yield Variates Presented in Table 17.1

Interval	Interval midpoint	Class boundaries	Tally
30–34	32	29.5–34.5	/
35–39	37	34.5–39.5	/////
40–44	42	39.5–44.5	///// ///// //
45–49	47	44.5–49.5	///// ///// ///// //
50–54	52	49.5–54.5	///// ///// ///// //
55–59	57	54.5–59.5	///// //
60–64	62	59.5–64.5	/
65–69	67	64.5–69.5	///
70–74	72	69.5–74.5	////
75–79	77	74.5–79.5	///// ///// /
80–84	82	79.5–84.5	//

Sometimes several different populations or **strata** (smaller subpopulations) might be used simultaneously. Equally important is the fact that when a population is defined, other populations are excluded. For example, populations involving soybean trials from other states, or even other soybean maturity groups, or even other Group V varieties available in Mississippi are not listed in Table 17.1. Always have a clear idea what population is being targeted when sampling.

Replication is another idea that needs to be discussed. When more than one individual of a population is uniquely subjected to a treatment in an experiment, it is said that treatment has been **replicated.** Often, replication is not easy to achieve and is abused in many experiments. Consider a trivial example. Suppose a food chemist wanted to determine the chocolate content of a candy bar. He realizes that he should replicate his study. So, he divides a single candy bar into four parts and calls each part a replicate and measures the chocolate content of each part. Has replication really been accomplished? As an answer consider the same experiment, but now the food chemist spends considerable *effort* to locate four candy bars from *different batches* produced by the candy maker. Each bar is now called a replicate, and the chocolate content of each bar is determined. It should be obvious that the second approach is much better than the first. A good rule of thumb (Milliken and Johnson, 1984, p. 50) to remember is that if replication is achieved by dividing (or splitting) a larger part into smaller parts then one has not properly replicated.

Replication is an important corollary concept to sampling because individuals, even though they come from the same parent population, are different. Each one will respond differently to the same treatment or differ from one another in some character that is being measured. Similarly, not all individuals in a sampled population are exactly the same. For example, the population of all 20-year-old males in Mississippi will not have the same height and/or weight. There will be **variation** in these characters. A statistic, called the sample **variance,** is a measure of the spread in how individuals in the population differ. The **average** response of sampled individuals is used to estimate (or **infer**) the center of the **distribution** of a population trait.

Two additional points are important for understanding how to sample properly. The first point is the fair or equal (**random**) opportunity for any individual in a population to be selected for measurement. This issue implies that there is no bias or prejudice in selecting or not selecting an individual during sampling. If you are uncomfortable with the word *random,* substitute the words *fair, representative,* or *unprejudiced* and you will have a good idea what statisticians have in mind when they use the word *random.* Always know the rules used to select the individuals included in the sample. You should know these even if someone else reports to you the results of a sample. The key point is that for every sample design there are rules that provide a basis for deciding which individuals are included in a sample. For example, if soybean yields for a 100-acre field are estimated using only a 1-acre tract that is the best (or worst) acre in the field, the answer obtained is not going to be of any real use. In this case, there was no rule established that gave the sampler a random sample (apart from a poor rule that said only to sample the same 1-acre tract of the field).

The second point is that the response or sample behavior of individuals can be predicted (within reasonable limits) again and again once that response has been observed for a particular set of conditions. As a result, smaller sample sizes from many similar populations can be pooled or collected together. It is not necessary to use large sample sizes with each population. The chief point is that large sample sizes alone do not result in an increased understanding of population behavior; the proper allocation of sampling effort is the principal key to estimating population traits successfully.

Sample data can be arranged into **one-** or **two-way tables**. A one-way table is a collection of data in which only one classification criterion exists. A two-way table is a collection of data with two classification criteria present. A classification criterion is any definition that can be used to identify a population. Classification variables are also called **independent variables**. The traits that are measured are not part of the classification criteria. The items measured are collectively called **dependent**, or **response variables**. To illustrate, Table 17.1 is a two-way table (ignore for the moment the average and standard deviation statistics) where *Variety* and *Location* are the two classification criteria for the response variable, yield. A one-way table of the data in Table 17.1 can be quickly made if variety is the only classification criterion and yield is again the response variable. Many other examples of one- or two-way tables could be given. Higher levels of data tables can also be constructed. The take-home message is that classification criteria define populations of interest.

It has already been mentioned that a good sampling plan establishes rules that specify (1) which population will be sampled and (2) the chances (i.e., **probability**) that a particular individual from that population will or will not be included in a sample (Thompson, 1992). Different sampling plans have different rules, which is the reason there are so many different sampling schemes. Specifically, the rules that determine if an individual is or is not included in a sample comprise the **randomization scheme** of a sampling plan. There are many different ways to create a randomization scheme. Often, this procedure is that part which gives a particular sampling plan its name. A few common names for different plans are simple random, stratified, sequential, cluster, two-stage, or adaptive sampling. Later, we will contrast some of the strengths and weaknesses of a few of these schemes.

The **sample units** that a particular sampling plan employs should be distinct, non-overlapping, and comprise the fundamental unit that is collected during sampling. Often, the sample unit will be a single individual plant or counts of individual insect pests at a specific stage of development. In this instance, the sample unit is defined by a natural or biological characteristic. Sometimes, the sampling unit will not be so discrete or well defined and will have to be defined by artificial characteristics established by the sampler. In these instances, the sampling unit may be arbitrary, such as a square yard or an acre of land. Even in this case, clear guidelines must exist to define the sample unit. Be wary of sampling plans that are vague in spelling out what is the basic sample unit. An important point raised by Ludwig and Reynolds (1988) is that a sample is a collection of sample units. It is incorrect to call a sample unit the sample.

Most sampling plans will have only one size of sample unit, but some can have more than one size. In some other sampling designs (e.g., belt transects or two-stage sampling designs), the sample unit size can vary. The **sample unit size** is related to questions or objectives established by the sampling plan. Some authors deal with this issue by labeling some sample units the primary units; other sample units used in the design are labeled secondary sample units.

The **sample number** (or **sample number size**) is how many sample units are collected or counted when a sampling plan is implemented. This number can be different for different fields or different sample dates. The sampler should always have in mind well-defined stopping rules that determine when enough sample units have been collected from a population.

Occasionally, a population is small enough that is easy to count completely — e.g., the number of coins of different denominations in the ashtray of a pickup. It would only take a few seconds to determine what was there and how much. Here, one can **exhaustively sample** the population. There is no need to estimate any population parameters for any population that has been completely counted. The answer is known without doubt and the sample size is the same as the number of members in the population.

More often, however, a population is large but countable if given enough time (e.g., the number of soybeans in a bushel). More likely, it is so large that it would be impossible or too expensive to count all the individuals (e.g., the exact number of soybeans stored in a grain bin). In agriculture, most populations of interest are of the latter type. Since this is true, the reason for developing and using a sampling plan should be obvious. Ruesink (1980) stated, "A 'sample' consists of a small collection drawn from a larger 'population' about which information is desired. It is the sample that is observed, but it is the population that is studied." Therefore, to develop a sampling design that can be used to study a large population, general principles have to be employed. When describing the principles that give a sampling plan its distinctiveness, most texts introduce mathematical notation. The use of notation is what makes the reading of sampling plans so difficult for many people. The style of mathematical notation used by different authors is often similar, but frequently different styles of notation exist. Therefore, it is easy to become confused by this unfortunate practice. We shall try to follow as much as possible the notation style of Little and Hills (1978). However, before this notation can be described, it is necessary to first introduce additional definitions.

Unique characteristics of interest measured from individuals sampled from a population without favoritism (or bias) are called **variables**. The individual observations that are recorded for a particular variable are called **variates** (Little and Hills, 1978). Examples of variables important to soybean production are yield/acre, number of plants/acre, average plant height, percent defoliation by soybean loopers, and so on.

Soon, data from Table 17.1 will be used to create artificially several different sets of samples drawn from different populations. These data are the yields of several (subjectively selected for no particular reason) Group V soybean varieties across several Mississippi locations during 1994 (Mississippi Agricultural and Forestry Experiment Station, MAFES, Information Bulletin 276). Calculations will be performed to estimate several statistics of interest from the variates that result for each "made-up" population. The examples that follow closely model a sampling plan known as **simple random sampling**.

In simple random sampling, each individual of the population has an equal chance of being included in the sample. This does not imply that every individual in the population *will be* sampled. Other sampling plans have different probabilities (or chances) than equal chances of including individuals. Different sampling plans offer different strengths and advantages. In fact, the data of Table 17.1 will eventually illustrate the need to use a stratified sampling plan rather than simple random sampling.

The most commonly used statistic is called the average (or **mean**) and is often represented by one of several symbols: $\bar{x}$ or $\bar{y}$ or $\hat{\mu}$. The average is one of three measures of central tendency. As the examples are worked through, the concept of what is meant by central tendency should become more obvious. For now, consider the following illustration. If two children of equal weight are simultaneously positioned on both sides of a teeter-totter, the two will balance and hang in midair. In this case, the distribution of weight across the board is symmetrical. Similarly, two children of different, but not dramatically different weights can be balanced by moving the position of each child to different distances from the middle of the teeter-totter. The average is the "point" on the teeter-totter board where the opposing forces are equal across the support bar.

Two other statistics of central tendency are known as the **mode** and the **median** and often are not reported. This is not to say that they are unimportant. For precision agriculture applications, they are more useful than the average because the variable of interest is frequently skewed toward large or small values.

Later, the concept of a histogram will be introduced. Soon thereafter, it will be seen that the average is the center of the histogram of a data set. If the histogram is perfectly symmetrical, the mode, median, and average (or mean) will be identical. If the distribution is **multimodal** (i.e, having multiple "peaks"), or **skewed** (i.e., **asymmetrical**), the three measures of central tendency for a set of sample data will differ considerably. Examples and graphs will be given later to illustrate different kinds of histograms and a statistical model will be fit to these distributions.

A second important statistic is called the sample **variance** (s^2). Another major statistic is the sample **standard deviation** (s) and is the square root of the sample variance. These last two statistics measure the spread of sample variates about the average.

Let us now examine what is meant by the terms *average, variance,* and *standard deviation* by pretending that the data contained in Table 17.1 can be used to represent different samples collected from several large imaginary populations. To do this exercise, several simple equations are introduced to summarize these sets of fabricated sample data using yield as the **response variable.** In the sampling literature, these equations are used to obtain estimates of the population average and variance.

The estimate of the population **average** for a **variable** is the sum of the **variates** (y_i) divided by the number of individuals sampled, or

$$\text{Average} = \frac{y_1 + y_2 + y_3 + \cdots + y_n}{n} = \frac{\sum y_i}{n}. \tag{17.1}$$

The symbol Σ means to "sum over" or, in this instance, take the sum over all the variates in the sample. This sum is next divided by the total number of observations (n). The subscripts ($i = 1, 2, 3, \ldots, n$) represent the first, second, third observations, along with other individuals included in the sample.

The estimate of the **variance** is a little more difficult but can (as the average) be performed with the help of a simple calculator. (See your instruction manual for the calculator or software package to find out if these capabilities are available.) Several formulas are available, but the one selected here is the following:

$$\text{Variance} = \frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n-1}. \tag{17.2}$$

To illustrate what Equations 17.1 and 17.2 involve, the average and the variance will be estimated across four locations for the Asgrow variety A5560 found in Table 17.1. Each step that follows corresponds to the different terms found in Equations 17.1 and 17.2. The average will be estimated first. To begin, correctly identify and list the variates for yield: 32.9, 50.8, 77.7, and 48.6 bu/acre. Next, obtain the **sum** of these four yields:

Step 1: Σy_i = sum = 32.9 + 50.8 + 77.7 + 48.6 = 210.0 bu/acre.

Next, divide the sum by number of variates (n = 4) to obtain the average:

Step 2: $\Sigma y_i / n$ = average = 210.0 ÷ 4 = 52.5 bu/acre.

To estimate the variance of this sample, multiply each variate by itself (known as taking the square of the observation) and obtain the sum of these values:

Step 3: Σy_i^2 = 12,062.3 = (32.9 × 32.9) + (50.8 × 50.8) + (77.7 × 77.7) + (48.6 × 48.6).

The sum obtained in step 1 is then multiplied by itself:

Step 4: $(\Sigma y_i)^2$ = sum squared = (210.0 × 210.0) = 44,100.

Divide the squared sum by the number of samples (n) and subtract this from the result obtained in step 3. You have now obtained the **numerator** of Equation 17.2 as shown below:

Step 5: $\Sigma y_i^2 — (\Sigma y_i)^2/n$ = numerator = 12,062.3 – (44,100 ÷ 4) = 1037.3.

Finally, divide the numerator by the **denominator** that is one less than the total number of variates ($n - 1 = 4 - 1$) or

Step 6: $[\Sigma y_i^2 — (\Sigma y_i)^2/n]/n - 1$ = variance = s^2 = 1037.3 ÷ (4 – 1) = 345.8.

Having now obtained the estimate of the sample variance, it is possible to determine the estimate (s) of the standard deviation, σ. Using a pocket calculator, take the **square root** of the variance:

Step 7: standard deviation = $s = \sqrt{345.8}$ = 18.6 bu/acre.

The units of measurement for the standard deviation are the same as that of the original observations (variates). For the variance, we have not associated units of measure or dimension. For example, with the data being used here, the units would be bushels squared per acre squared, or $\text{bu}^2/\text{acre}^2$. These are difficult units of measure to interpret and often the variance is reported as a dimensionless number. Remember that these values are estimates of parameters and are not the true but unknown population values (Freese, 1967).

The smaller the variance (or standard deviation), the closer all variates are to the average. Later, several histograms (e.g., Figures 17.8 and 17.9) will be pictured for sets of samples obtained from different populations. Examine these graphs closely to see how different values of variates influence the estimates of the average and variance for the different populations. With practice, you will gain an intuitive understanding for what these statistics are telling you for sets of data acquired from your farm.

There is a shortcut method (Miller and Freund, 1977) that exists to estimate the standard deviation for small-sized samples (two to ten observations). The method has one assumption. It assumes that samples are drawn from a population whose character of interest follows a **normal distribution**. The probability plotting method discussed later in

Table 17.2 Sample Size and Constants Used to Estimate the Standard Deviation for Small Sample Sizes

n	2	3	4	5	6	7	8	9	10
d_2	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970	3.078

this chapter can be used to confirm if a set of data is from a normally distributed population. The normal distribution is a bell-shaped curve that is symmetrical about the average. For symmetrical distributions, the average, mode, and median are equivalent, whereas for asymmetrical (skewed) distributions this is not the case.

This shortcut method uses the **range** (R) of the sample to estimate the standard deviation. The range is the difference between the smallest and largest variates included in a sample. The range is divided by a special number, labeled in this discussion as d_2 (found in Table 17.2) to estimate the sample standard deviation, or $s = R/d_2$. The first row of Table 17.2 depicts the sample size from two to ten sampling units, whereas the second row presents the constant (d_2) used for that size of sample.

Using the data of Table 17.1, several examples now follow to illustrate the use of Table 17.2 to estimate the standard deviation as compared with taking the square root of the variance obtained in Equation 17.2. (Note that the estimates reported in Table 17.1 for the standard deviation were obtained with a software package that used Equation 17.2.) For practice, you may wish to calculate and compare additional values. The shortcut method to estimate the standard deviation (s) for the Asgrow brands (A5560, A5843, A5885, and A5979) at the Clarksdale location uses the information of both Tables 17.1 and 17.2 as follows:

$$s = (54.0 - 32.9)/d_2 = 21.1/2.059 = 10.2 \text{ bu/acre}.$$

The standard deviation for variety A5885 across the four locations (Clarksdale, Rolling Fork, Hernando, and Verona) is estimated as follows:

$$s = (75.8 - 40.9)/d_2 = 34.9/2.059 = 16.9 \text{ bu/acre}.$$

The previous examples all involved sample sizes of $n = 4$. One last example will use the Hartz brand varieties ($n = 8$) at the Rolling Fork location. Here the estimate of the standard deviation is as follows:

$$s = (57.0 - 42.0)/\mathrm{d}_2 = 15.0/2.847 = 5.3 \text{ bu/acre}.$$

Additional examples could be given throughout the table, but with the few that have been done it is easy to see that the agreement is within 1 to 3 bu/acre at most, and that the estimates were obtained with less effort. We think this is a very practical way to estimate the standard deviation for most samples having two to ten observations. If more observations are present then one should use a calculator or software package to obtain the estimate.

The examples just completed involved small numbers of samples. In actual practice, an important question to consider is how many sample units (i.e., the sample size, n) are needed to acquire suitable information to make a management decision? The largest sample size that time and effort allows is not always the best number to collect. A balance must be reached between improving the precision of the sample and increasing the cost of sampling while experiencing a diminishing return on learning new information as the sample size increases. But, if the sample size is too small, then poorer or even useless estimates of population traits result. Generally, the more variation about the average, the larger the sample should be to characterize the population, or the sample should be stratified if rules to develop strata can be devised. Presently, other rules will be given as we focus our attention on determining the proper number of samples.

One important consideration that influences the sample size (n) is the distribution of the character being sampled. The shape or pattern the sample data portray can be inferred from a **histogram** (see below). It has been found that agricultural data, when represented by adequate sample sizes, commonly display a histogram that can be described by a special equation known as the **normal frequency distribution**. At times, the histogram may suggest another distribution to use that is not the normal distribution (D'Agostino and Stephens, 1986). When samples are not normally distributed, one has to either (1) find a sampling plan that accommodates the non-normal behavior of the sample data or (2) make modifications to the normal distribution so that it can still be used. The details of how to do either of these alternatives are beyond the scope of this chapter. Other texts like Thompson (1992) should be consulted. The message is that sample data can follow different sampling distributions. The chief concern is to be aware that data may be symmetrical or asymmetrical about the estimate of the average. We provide here several tools to identify key behaviors of the sample data.

The equation that models or approximates the shape (or histogram) of the normal frequency distribution has the following form (Little and Hills, 1978):

$$\text{Normal Frequency Distribution} = f(y) = \frac{N}{\sigma\sqrt{(6.291)}}\,e^{-(y-\mu)^2/(2\sigma^2)}. \qquad (17.3)$$

There are several parts to Equation 17.3 that should be pointed out. First, the value N is the total number of individuals in the sample for a specific character. The number N can be very large, but it is a finite number that is countable. The values μ and σ are the parameters for the average and the standard deviation (where if it were possible to count all individuals in the population they would be numbers whose values are known without doubt). The estimate of μ is the sample average ($\bar{x}$) and σ is estimated by the sample standard deviation s, which is the square root of the estimated sample variance, s^2. The value 6.291 (or 2 * 3.1456 = 2 * pi, or that famous value from the geometry of a circle) is a special constant determined to be necessary by experience.

The formula (Equation 17.3) can also help understand why certain requirements and assumptions are specified for a sampling plan. The parameter σ shown in Equation 17.3 has a large influence upon deciding the sample size or number of samples to take. Generally, as σ gets larger, the sample size must be larger to characterize a population at a given level of precision. There are exceptions to this rule of thumb. For example, using a sampling plan that stratifies the population into smaller strata where the variance is different for each strata can help preserve the use of small samples. Examine closely the material you may acquire about a particular sampling plan to see how it addresses the issues about sample size.

Another thought of interest about Equation 17.3 is to consider how multimodal distributions influence the estimate of the average and standard deviation. Sometimes the population to be sampled will have a bimodal or perhaps a multimodal distribution. A bimodal distribution has two peaks (called **maxima**); a multimodal distribution has several peaks. These peaks represent variates that are most common in a set of sample data. The peaks do not necessarily have to be of the same size, or if more than one, the same distance apart from one another. Naturally, a unimodal distribution will have only one peak.

Table 17.3 Example Sample Data of Soybean Plant Heights (in.) Categorized into Irrigated (n = 50) and Nonirrigated (n = 50) Portions of a Field (See text for further discussion)

Irrigated plant heights		Nonirrigated plant heights	
20.4	16.0	20.4	13.8
19.6	20.1	16.2	16.3
19.9	25.2	14.0	14.8
21.5	17.5	12.7	14.9
17.4	22.6	14.1	9.5
20.0	17.1	7.7	17.4
21.2	15.1	17.6	16.5
22.4	22.4	14.0	14.5
18.7	20.4	13.5	11.0
19.1	15.1	14.8	13.7
20.5	24.4	12.8	13.8
22.4	18.9	9.2	18.8
21.8	21.5	15.1	12.5
20.1	18.7	14.4	13.5
18.2	19.4	17.0	12.3
18.0	21.4	13.9	13.8
20.8	22.7	15.7	13.5
19.3	21.7	19.9	12.9
13.8	17.5	15.7	11.3
20.6	20.6	12.4	15.2
21.0	19.7	19.5	15.2
20.3	19.4	16.6	15.1
24.2	19.3	15.0	16.4
18.9	16.7	17.6	14.2
19.5	21.5	20.5	15.5

How does the occurrence of multiple peaks in a collection of sample data influence the use of a sampling plan? Several answers are possible, but one important possibility can be illustrated by a simple example. Suppose the objective is to estimate the height of soybean plants from a particular field where only 80% of the acreage is irrigated. Lately, it has been dry so irrigations have been applied. The sampler collected 100 samples (Table 17.3), with half of the samples selected from the irrigated portion and half from the nonirrigated portion. The frequency distribution of plant heights for this population of soybeans can be expected to follow a bimodal distribution due to the influence of irrigation upon plant growth.

Our first question to the reader is what has the sampler overlooked, or forgotten to do? A second question to the reader is what will be the characteristics of the estimates of the average and the variance? Make a guess to both questions before reading further and, if you wish, apply Equations 17.1 and 17.2, using the variates found in Table 17.3.

Without stratifying (How many populations are being sampled?) the sample between irrigated and nonirrigated plants, the estimate of the population average and variance will behave as if the data were taken from a unimodal population when in fact the actual sample data are bimodal (See Figure 17.1). In this instance, the estimate of the average for the 100 plants is 17.3 in. with a standard deviation of 3.6 in. The curve drawn behind the bars is the shape of the **normal probability density function** (see below) using the above estimates for the population parameters μ and σ.

The sampler overlooked the fact that irrigation can result in the plants having different heights in the two parts of the field. The sampler in this instance made the mistake of not

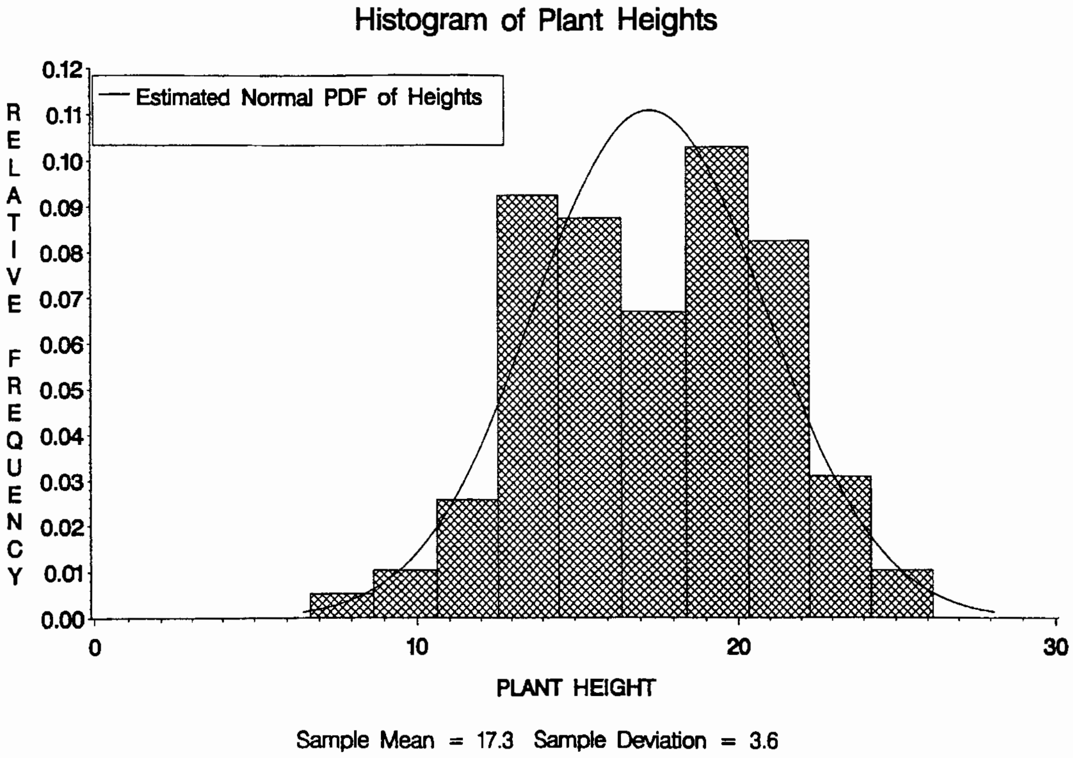


Figure 17.1 Histogram of the soybean heights found in Table 17.3. The histogram clearly shows the bimodal behavior of these data. The line drawn behind the bars is the fit of the normal pdf to data and ignores the bimodal trait of the data. The frequencies of each class (the bars) are expressed as relative (not absolute, see text) values on the *y*-axis to the right. The base of each bar is 2 units (in.) of plant height; therefore, sampled plants that are within 2 in. of each other in height are placed into the same class.

stratifying the sample population. The sampler should have divided (i.e., have used a stratified sampling plan) the field into two populations (irrigated and nonirrigated) and then sampled and measured the height of randomly selected plants in each part. If this had been done, the following results would have obtained. The average and standard deviation for the 50 plant heights of the irrigated part of the field is 19.9 in. and 2.4 in. The nonirrigated field plant height average was 14.8 in. with an estimated standard deviation of 2.7 in. Notice that precision was improved by stratification since both estimates of the standard deviation were smaller than the estimate obtained without stratification. The consequence of treating two separate populations as one depends upon the goals established before the samples were collected. At times, serious implications could result; at other times, the effect will be moot. You should keep in mind that sampling plans (e.g., stratified) exist to handle bi- or multimodal distributions of sampled characteristics.

A chief feature of a **stratified sampling** plan is that the sampler is making use of knowledge about the population to do a better job of sampling the population. Stratified sampling divides the population into subpopulations, or strata, that are less variable than the original population (Cochran, 1956). The strata do not need to be the same size in terms of area or number of individuals. When a population has been stratified, different parts of the population can be sampled at different rates (i.e., the number of samples can be different for different strata). This means that sampling effort and resources can be efficiently allocated to those areas where more samples are necessary to characterize the population. The sample data are collected and the parameters are estimated for each strata. The estimates of the strata can later be combined (appropriately) to estimate population parameters with better precision than a simple random sampling plan. The appeal of a stratified sampling scheme is to reduce the **error of the sample**. The previous example of soybean plant heights demonstrated this fact. The error (or **bias**) of a sample is the difference between the estimate and the true (parametric) population value.

The ability to tell if a population should be stratified into different parts can be a good skill to acquire. Practically, whenever a population is stratified, it is an admission that one part of the crop in a field might be managed differently than other parts. In fact, this concept is what lies behind the recent emphasis upon precision agriculture. More comments to this issue from the perspective of sampling for soil characteristics will be given later in this chapter. The failure to use a stratified sampling plan when necessary can lead to wasted resources.

Another common sampling plan is one called **systematic sampling** (or systematic sampling with a random start). Concepts of this sampling plan can be incorporated into other sampling designs. The chief advantage of this plan is reducing the time in selecting sample units. If all samples are randomly drawn, considerable time can be involved. A simple example follows. Suppose one wanted to estimate the average number of soybean pods per plant in a uniform field of 35 acres. The basic sample unit is three successive individual plants. The sample character to be measured is the number of pods per plant. The desired sample size is that at least 35 sets of three plants are to be sampled throughout the entire field (at one set per acre). The sampler has determined that she can walk into a new acre about every 156 paces along a series of parallel lines placed 37 rows apart. Starting at the Northeast corner of the field, she chooses from a random number table two numbers (random number tables can be found in statistics books or generated by a calculator or software package on a computer). The first is a number less than 37, say, 17, and the second is a number less than 156, say, 97. These two numbers define the placement of the first line and the paces on this line to locate the first sample of three plants. For a different sampling date, she will select two new numbers.

With these numbers, she moves to the 17th row from the corner and then walks 97 paces into the field. She will count the number of pods on three successive plants closest to her left foot when she reaches the 97th pace. Thereafter, she will collect samples every 156 paces and over 37 rows until she reaches the opposite side of the field. Paces left uncounted at the end of a line will be resumed on the next line until 156 paces have been walked (idealistically, the line is one long unbroken line). The pattern is repeated until the end of the field is reached. The starting point was selected at random, but the samples thereafter use the same distances of 37 rows between lines and 156 paces apart on each line. The characteristic of systematic sample plans is that the choice of the first member of the sample determines the whole sample (Cochran, 1956). Another advantage is that, with planning, the samples can be more evenly distributed over the defined population. The chief disadvantage is that the samples may follow (by chance) some periodic variation present in the population (Cochran, 1956) and introduce a source of bias. Thus, it may be necessary to stratify the field and use a systematic sampling plan in each strata.

Some sample variates, as mentioned earlier, follow a **skewed** distribution (See Figure 17.13) and are not symmetrically distributed about the mode. For skewed data, the median, mode, and average will be more dissimilar. These distributions can be skewed to the left or to the right. It is suspected that many soil characteristics, when considered on a field-level basis, will follow a skewed distribution. The best sampling plan to use for skewed variates is a modified form of stratified sampling in which the number of samples allocated to a strata is **disproportionate** to the size of the strata (Cochran, 1956). Returning to our plant height example, one could expect the height of plants to vary more in the nonirrigated part than in the irrigated part. So, perhaps it should be considered to have a larger sample size in the nonirrigated than the irrigated portion. Thus, with a stratified sampling plan, the requirement that each individual have an equal chance of being included in the sample refers only to sample units within a stratum and not to the fact that sample units in every strata have the same chance of being sampled as units found in another stratum. The challenging task is to determine the boundaries of the strata followed next by determining the appropriate sample size to use in each stratum. Several years of effort and information from several sources may ultimately have to be used to define the boundaries of strata properly. Thus, in the first year of some sample studies, use the greatest effort and resources on only a small part of a field. Learn a little about a small part, but at a high level of quality. As more is learned, expand the scope of the sampling effort.

At times the distinctions between multimodal and skewed distributions of population attributes become blurred and have little impact. On other occasions, these possible distributions are quite different from one another and can introduce considerable error into sample estimates if not taken into account. The key point is to be aware of these possibilities for any set of sample data. If these features are present, be practical in taking them into account when developing management policy about the crop. The technique of **probability plotting** will be discussed later as a tool to determine if samples are multimodal or heavily skewed.

Sample data can be **discrete** or **continuous**. Discrete data are usually whole numbers. Continuous data are fractional or integer values that occur continuously between specified limits (Fogiel, 1985). Special equations called **probability functions** mathematically describe different behaviors of sample data (Gelman et al., 1995). If the sample data are continuous, the distribution is called a **probability density function**. If the data are discrete, the distribution is called a **probability mass function**. The graph of discrete data looks like a series of small stair steps. Probability functions possess two main features. The area under the curve sums to one and the observed probability values associated with variates of a variable are between 0 and 1 (or 0 and 100% if probability is expressed as percentages). Therefore, one can never have a probability of 2 (or 200%) or –1 (or –100%). However, one can observe a 100% or 200% increase or change in the rate at which some process proceeds. Do not confuse rates of change with the probability of occurrence of some action or item of interest.

The normal frequency distribution has been described in Equation 17.3. The **normal probability density function** is

$$f(y) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-(y-\mu)^2/(2\sigma^2)}. \tag{17.4}$$

The form of Equation 17.4 is very similar to Equation 17.3. The difference between the two is that N has been replaced by 1 with the result that Equation 17.4 now describes areas under a symmetrical bell-shaped curve whose total area is 1. Before, Equation 17.3 modeled the frequency of sampled individuals in different class intervals. Equation 17.4 meets the requirement that the probabilities of occurrence for the different variates (y) in the sample will be between 0 and 1 and that the sum of all the probabilities will be exactly 1. Sometimes the normal probability density function is standardized to have a mean of 0 and a variance of 1. If this is the case, it is called the **standard normal density function**. Sample data that are normally distributed can be transformed to the standard normal function by subtracting the sample average from each of the variates and dividing each difference by the estimate of the standard deviation. These values are called z-scores or standardized scores (Gonick and Smith, 1993) and range between –2 and 2 for 95% of the time. A few z-score values will fall in the intervals –4 and –2 and 2 and 4. These scores are not probabilities and do not have to be restricted between the values 0 and 1.

Probability is a difficult concept for many people. One simple idea to keep in mind is that probability can be a measure of the strength or value of a proposition or occurrence of an event. A probability of 0 means that the event will never occur or that a proposition is not very believable. A probability of 1 means that the event will most definitely occur or that the proposal is very believable. A probability of 0.5 means that an event will occur about one half of the time or that the degree of belief about a proposition is near neutral. Another concept of what is meant by probability is to consider that it is the long-term relative frequency of the occurrence of an event or outcome of an experiment. For example, a fair coin when tossed will turn up as "tails" about 500 times in 1000 tosses for a probability of 0.5.

A good book to help you assimilate and improve your skills is the text by Gonick and Smith (1993). This inexpensive book uses small, related cartoons and simple diagrams to discuss data manipulation, statistics, and sampling. It can be read through in a week or so with 1 or 2 hours of effort per night. Those readers who are interested in precision agriculture, but are unfamiliar with the mathematical concepts that buttress this technology, should consider reading this text.

Determining the size of a sample

Determining the appropriate number of units to sample is challenging. Far too often it is never done. The challenge of modern sampling plans is to balance the **precision** of the sample (or how close the estimate is to the true but unknown population parameter) against the **cost** of taking the sample. For example, if the sample is too small, the estimate will be too inaccurate to be useful (Cochran, 1956). On the other hand, if the sample is too large, it will cost too much in time and labor to collect the samples. Plainly stated, "Samples cost money. So do errors. The aim in planning a (sample) should be to take enough observations to obtain the desired precision — no more, no less" (Freese, 1967). Thus, the proper sample size is a balance between the forces of having the sample be informative and not being too expensive.

The rule of thumb or formula we present to determine sample size is the "confidence limit" approach described by Cochran (1956). Assume that the sampler is willing to take a 5% chance that the allowable error will exceed some value, say *L*. The following formula to estimate the size of sample *n* can be derived as

$$n = \frac{4\sigma^2}{L^2}. \tag{17.5}$$

To use Equation 17.5 one must have an estimate of σ from previous samples or knowledge. A simple example is to return to the data of plant heights in Table 17.3. For irrigated soybean plants, how many plants should be sampled if it is desired to estimate the true plant height within 1 in. with a 5% risk that the error will exceed 1 in.? Earlier, it was estimated that the standard deviation for irrigated soybean plant heights in this field was $s = 2.37$. By making use of this estimate in Equation 17.5, the following estimate of sample size is obtained:

$$n = \frac{4 \cdot (2.37)^2}{(1)^2} = 22.47 \approx 22 \text{ plants}.$$

Thus, if 50 samples were actually taken (as shown in Table 17.3), then 28 more plants than necessary were sampled. Cost, in this case, could be considerably reduced by taking only 22 samples instead of the 50 plants illustrated in the example.

For samples that have large standard deviations, or when the sampling objective demands high quality, the estimates of the sample size can be quite large. For example, to be within ½ in. of estimating plant heights for the previous example, a sample size of about 90 plants is required. Here, at this level of precision not enough plants are included in a possible sample. Clearly, sample size is influenced by variability and the degree of precision required with the estimate.

If the sampled attribute is a **binomial proportion**, the following formula should be applied:

$$n = \frac{4pq}{L^2}. \tag{17.6}$$

A binomial proportion is a ratio between two quantities — the number of individuals classed "yes" and the total number of individuals examined for a specific character (Cochran, 1956). Typically, in row crop agriculture the characteristic of interest is the proportion, or percent, of plants infested with insects or disease. The values p and q are related by the expression $p + q = 1$ if they are proportions, or $p + q = 100$ if they are percentages. The values for p and q in Equation 17.6 should be expressed in the same units.

If Equations 17.5 or 17.6 return estimates that are more than 10% of the total population size, the following correction can be applied:

$$n' = \frac{n}{1+\phi}. \tag{17.7}$$

The quantity ϕ is the ratio between the total number of individuals that could be sampled and an initial estimate of the sample size that was too large to be practical. To illustrate the application of Equations 17.6 and 17.7, we present the following example of Cochran (1956). A cursory inspection of 480 seedlings suggests that about 15% are diseased. What size of sample is needed to determine p, the percent that are diseased, to within ±5%? By Equation 17.6, the following result is obtained:

$$n = \frac{4\,(15)\,(85)}{(5)^2} = 204 \text{ seedlings}.$$

This estimate of sample size is almost half of the total, and with a sample this large one may just as well examine the whole batch. However, a revised sample number can be obtained from Equation 17.7 that would help avoid the need to examine the whole group. The revised number is

$$n' = \frac{204}{1+\dfrac{204}{480}} = 143 \text{ seedlings}.$$

Remember, we applied Equation 17.7 since the initial estimate of n was more than 10% of the total of the original population. Other authors present other formulas, but the expressions provided above by Cochran (1956) are not as cumbersome to apply. The expressions provided here apply only to simple random sampling plans. More elaborate sampling plans should report applicable formulas (e.g., see Thompson, 1992), but the rules of thumb provided here are good starting points for other sampling plans.

Graphical techniques

Graphical techniques for analyzing data have been available for centuries. However, the advent of computers with greater speed, smaller size, and affordable pricing has caused dramatic changes in graphical methods. Numerous methods of representing data exist or are being developed (e.g., Gazey and Staley, 1986; Friendly, 1991; Fortner, 1995). In this section, we do not describe in great detail any particular graphing technique. However, several pencil and paper methods applicable to small-sized data sets are described in some detail. One needs to understand and appreciate these simpler methods to interpret properly and avoid abusing the power of modern graphical software packages.

Interesting studies exist that demonstrate how one could abuse computer or data management methods (e.g., Fortner, 1995; Calvert and Ma, 1996). Fortner (1995) provides an excellent discussion on how to organize and process data, and avoid common pitfalls that occur while working with technical data (many of the points that are raised apply to sample data as well). Any producer or consultant who works with data and computers should read this book.

Frequency tallies

A **frequency tally**, or frequency distribution table, is a simple but useful tool that summarizes sample data. It is a preparatory step to building a graph. The first step in constructing a frequency tally is to determine suitable **class intervals.** A class interval is the numerical distance, or span, between two whole-number end points that are called the **class limits.** The size of a class, or the class interval, is the difference between its upper and lower class limits. These intervals should be of equal length with midpoints that are convenient whole numbers (Gonick and Smith, 1993). Thus, a class will be a collection of sampled individuals who share similar values for a sample attribute.

The establishment of the appropriate number of classes for a set of sample data is important. Doing this is not an exact science and requires judgment on creating a balance between having too many classes and too few. The number of classes to use generally depends on the number of observations and their range (Little and Hills, 1978). A good general rule is never to use fewer than 5 or more than 15 or 20 classes (Miller and Freund, 1977). Also, a histogram should be constructed from at least 20 or more variates. Other important attributes of a frequency distribution are that its classes will not overlap and be the same size, it will accommodate all the data, and that its class limits will have the same number of decimal placeholders as the original data (Miller and Freund, 1977).

The **class boundaries** can be obtained by adding the upper limit of one class interval to the lower limit of the next higher class interval and dividing by 2 (Spiegel, 1962). If a variate has the same value as a class boundary, a general rule to follow is always to place these variates into the next higher class. The number of variates belonging to each class is called the **class frequency** (Spiegel, 1962).

The data of Table 17.1 will be used to illustrate the making of a frequency tally. It is easier to tally the data if they are first sorted in order from the smallest to the largest value. (The data of Table 17.1 have been sorted in Table 17.4 for use in building a probability plot, a technique that will be discussed shortly.) The smallest variate for yield is the variety A5560 grown at the Clarksdale location (32.9 bu/acre). The largest variate is for

Table 17.4 Rank Order Numbers and Probability Plotting Positions of Soybean Yields (sorted here in ascending order, found in Table 17.1)

Rank-order number	Variety	Yield	Plotting postion	%	Rank-order number	Variety	Yield	Plotting postion	%
1	A5560	32.9	0.123	1.23	41	H5454	51.1	0.506	50.62
2	9501	36.8	0.025	2.46	42	9593	51.7	0.518	51.85
3	9551	37.1	0.037	3.70	43	H5088	51.8	0.531	53.08
4	9501	37.5	0.049	4.94	44	H5545	52.3	0.543	54.32
5	H5350	37.8	0.062	6.17	45	9592	52.9	0.555	55.55
6	9551	38.9	0.074	7.41	46	9593	52.9	0.567	56.79
7	A5885	40.9	0.086	8.64	47	9501	53.1	0.580	58.02
8	H5566	42.0	0.099	9.88	48	9592	53.1	0.593	59.26
9	H5810	42.0	0.111	11.11	49	A5885	53.5	0.605	60.49
10	DP3589	42.1	0.123	12.34	50	9584	53.8	0.617	61.72
11	H5810	42.9	0.135	13.58	51	A5979	54.0	0.630	62.96
12	H5545	43.1	0.148	14.81	52	H5545	54.2	0.642	64.20
13	DPL105	43.3	0.160	16.05	53	DPL415	54.6	0.654	65.43
14	A5843	43.4	0.173	17.28	54	H5088	55.4	0.667	66.67
15	H5566	43.6	0.185	18.52	55	A5979	56.6	0.679	67.90
16	A5843	44.0	0.197	19.75	56	H5454	57.0	0.691	69.13
17	H5350	44.0	0.209	20.99	57	A5885	58.2	0.704	70.37
18	H5810	44.2	0.222	22.22	58	DP3589	58.4	0.716	71.60
19	Hartz516	44.6	0.234	23.46	59	A5843	58.5	0.728	72.83
20	9584	44.9	0.247	24.69	60	DP3589	62.6	0.740	74.07
21	A5979	45.2	0.259	25.93	61	9501	66.6	0.753	75.30
22	H5088	45.7	0.272	27.16	62	H5810	67.4	0.765	76.54
23	DPL105	46.2	0.284	28.36	63	9551	68.1	0.777	77.78
24	Hartz516	46.3	0.296	29.63	64	DP3589	69.5	0.790	79.01
25	H5218	46.7	0.308	30.86	65	H5545	72.6	0.802	80.24
26	H5218	47.5	0.321	32.10	66	H5088	72.8	0.815	81.48
27	DPL415	48.1	0.333	33.33	67	DPL415	73.5	0.827	82.72
28	H5566	48.2	0.345	34.57	68	H5566	74.9	0.839	83.95
29	DPL415	48.4	0.358	35.80	69	Hartz516	75.5	0.851	85.18
30	A5560	48.6	0.370	37.04	70	A5885	75.8	0.864	86.42
31	H5350	48.9	0.383	38.27	71	9593	77.0	0.876	87.65
32	H5218	49.1	0.395	39.51	72	H5350	77.1	0.889	88.89
33	H5454	49.3	0.407	40.74	73	A5560	77.7	0.901	90.12
34	Hartz516	49.3	0.419	41.97	74	A5843	77.7	0.914	91.36
35	9551	49.4	0.432	43.21	75	H5454	77.9	0.926	92.59
36	9593	50.5	0.444	44.44	76	A5979	78.2	0.938	93.83
37	A5560	50.8	0.457	45.67	77	DPL105	78.6	0.951	95.06
38	9584	50.9	0.469	46.91	78	H5218	79.4	0.963	96.30
39	9592	51.1	0.481	48.15	79	9584	81.0	0.975	97.53
40	DPL105	51.1	0.494	49.38	80	9592	84.7	0.988	98.76

The yields are plotted on the *y*-axis and the plotting position as percents are plotted along the *x*-axis of Figure 17.10.

variety 9592 grown at the Hernando location (84.7 bu/acre). The difference between the smallest and largest value (the **range**) is 51.8 bu/acre. By using the range, and a decision to declare that yields that differ by more than 5 bu/acre are different, it was determined that 11 classes should be constructed. This number was reached by merely dividing the range by 5 bu/acre and rounding up to the next whole number. The lowest interval was chosen to begin at 30 bu/acre. Table 17.1A shows the intervals, interval mid-points, class boundaries, and tally marks of each class. The interpretation of these data will be provided as the example is discussed.

Notice first of all that the tally marks create a crude histogram (turn Table 17.1A on its side by rotating it to the left). In brief, the data are bimodal (i.e., there are two peaks), with one peak occurring near the two midpoints of 47 and 52 bu/acre, and another smaller peak occurring at the class midpoint of 77 bu/acre. These facts suggest that Table 17.1 reflects several different patterns or processes at work that determine the yield response of these 20 varieties across the four locations. We next learn how to construct a histogram.

Histograms

Constructing a bar graph known as a histogram uses ideas similar to those used to build a frequency distribution. However, now the goal is to draw a more formal picture of the data. The histograms presented here were drawn using a graphical plotting package (Proc GCHART, SAS Institute, 1990). Many spreadsheet packages can also draw histograms.

Unlike the crude histogram shown in Table 17.1A, histograms can have the feature that the base of the bar multiplied by its height results in an "area" that is the class frequency. The base of these bars are centered on the **midpoint** (Gonick and Smith, 1993). Notice that in Figure 17.1, the base of each bar corresponds to a class interval of size 2. In Figures 17.3 through 17.7 the class interval is 5, but for Figure 17.2 the base of the bar is idealistically one and each bar is centered over a midpoint. Note that the frequencies on the *y*-axes (or vertical scale) are expressed as **relative frequencies** (Gonick and Smith, 1993) in Figures 17.1 and 17.3 through 17.7. In Figures 17.2, 17.8, and 17.9 the absolute frequencies are used; therefore, the area of each bar is the class frequency. For histograms that graph relative frequencies (like Figures 17.1 and 17.3 through 17.7), the area of each bar must be multiplied by the sample size (the total number of observations used in the sample) to get the original frequency of each class.

By comparing Figure 17.2 and Figure 17.7, it can be seen that either graph gives a similar picture of the data even though the two figures use different units on the *y*-axis. By selecting different graphing options, both figures could be drawn with the same labels on the horizontal scale or *x*-axis. A careful look shows that both Figures 17.2 and 17.7 have the same midpoints. This comparison illustrates that you should be alert when examining a histogram and notice which form (absolute vs. relative) is used to present the data on the vertical scale.

Examining the histograms of different samples provides a better understanding of how the average, standard deviation, and variance of different populations compare. For example, Figures 17.3 through 17.6 present both the group (dotted line behind the bars) and the subgroup (solid line behind the bars) "fits" of the normal probability density function (pdf). The "group" fit is the yields of all 20 varieties over all locations, while the "subgroup" fit is the yield of the 20 varieties at each location. Overall, only the Rolling Fork data is the most "normal," while the Clarksdale and Hernando locations are "okay," but slightly skewed. The Verona population appears to be the "poorest" fit to a normal curve because its largest frequency class lies to the right of the fitted mode.

Similarly, Figure 17.7 compares the fit of the normal pdf (solid line behind the bars) of the entire group, but overlays the normal density function fit to only the Hernando location (dashed line). Here, it is seen that the fit of the normal pdf to the entire set of yields is poor because it does not capture the bimodal pattern revealed by the bars of the histogram. Overlaying on the same plot the trace of the normal probability function fit to only the Hernando location (compare Figure 17.5) brings further emphasis to the fact that more than one statistical model is needed to describe these data properly. The dashed line clearly shows which bars of the combined data contain the yields from the Hernando

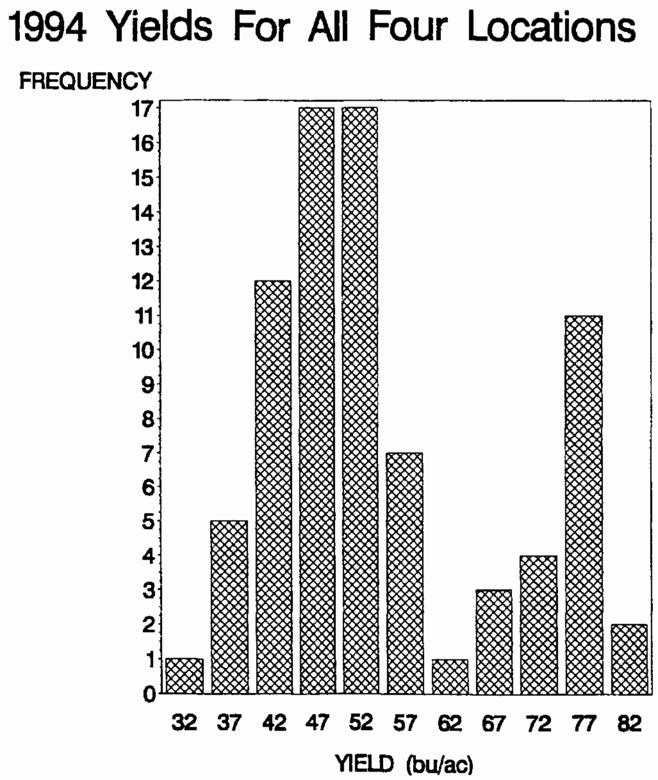
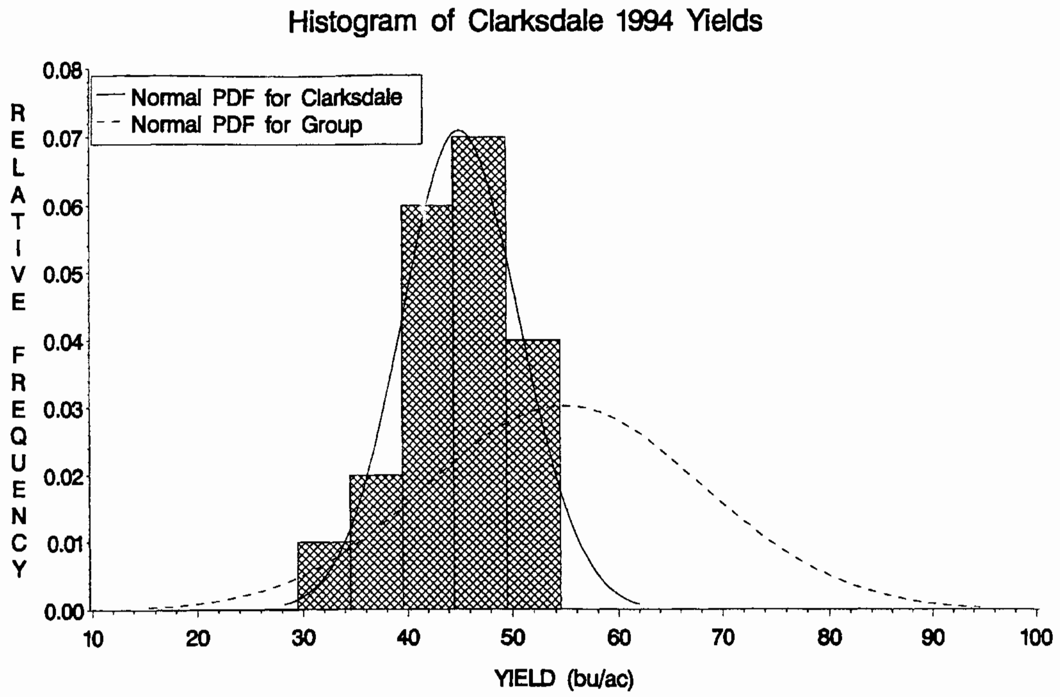


Figure 17.2 Histogram of absolute frequencies of soybean yields from Table 17.1 in 11 classes having class widths of 5 bu/acre. The y-axis of this histogram displays the actual (or absolute), not the relative, frequencies of individuals in each class.

variety trial. Interestingly, the Clarksdale, Rolling Fork, and Verona yield data collectively form a population that can be modeled quite well by the normal density function as judged by the shape of the remaining bars that lie to the left of the Hernando trace (Figure 17.7).

The consequences of a poor "fit" depend upon the sampling objectives. You, the decision maker, have the responsibility for determining whether or not these consequences matter for sample data under your control. For example, several histograms for the data of Table 17.1 strongly suggest that the Hernando data should be considered as a distinct population or strata.

Two, three-panel histograms have also been drawn to illustrate further how estimates of an average and a variance relate to each other in describing sample data. In Figure 17.8, the variance is the same for each panel, but the estimate of the average (mean) differs by 10 for values between 30 and 50. The shape of each distribution is the same, but its location on the x-axis is different. In Figure 17.9, the means (or the center of the distributions) of three populations are the same at a value of 50, but the variances differ considerably over values of 5, 25, and 50. Here, the location of the distribution is the same on the x-axis, but the shape varies from narrow to broad. In Figures 17.8 and 17.9, unlike Figures 17.1, 17.6, 17.7, and 17.13, the distributions are symmetrical; thus, estimates of the average, median, and mode will be equivalent statistics. As you work with more data sets, you will learn of other behaviors. The one pattern you should keep in mind is the fact that sample data expressed as a percentage or proportion will become more skewed (to the right) as the mode gets closer to zero.



Clarksdale Mean = 45.1 Clarksdale Std dev = 5.6
Group Mean = 55 Group Std dev = 13.2

Figure 17.3 Histogram of relative frequencies of soybean yields for 20 varieties in 1994 at the Clarksdale location. The subgroup mean and standard deviation are for only the yields obtained at the Clarksdale location; the group mean and standard deviation pertain to all the yields found in Table 17.1.

Probability plots

The following discussion is based on information from King (1980) and D'Agostino and Stephens (1986). A probability plot is constructed on **probability paper** by graphing the calculated plotting positions (x_i) of the sample variates (y_i) arranged in **rank order**. Data are rank-ordered when they are sorted from the smallest to the largest value and each variate is assigned a rank.

Probability paper is a special type of graph paper. For the soybean data of Table 17.1, yields comprise the ranked y-values on the **ordinate** or vertical axis. The x-values (or the probability plotting positions) are arranged by increasing value on the **abscissa** or horizontal axis. The plotting positions and ranked values of the variates comprise **ordered pairs** (x,y). Probability graph paper can be obtained from TEAM (Box 25, Tamworth, NH 03886; telephone 603-323-8843). Request a catalog, or ask for pricing and quantities on No. 3111 graph paper for sample data sets having 100 or fewer points, or No. 3211 graph paper for data sets having more than 100, but less than 10,000 values. Software can also be carefully programmed (e.g., a spreadsheet) to accomplish the same task as using special graphing paper.

The value for x (or probability plotting position) is given by a simple formula, or rule. The ith plotting position can be determined by $x_i = i/(n + 1)$, where i is the rank number and n is the total number of values in the set of data. The plotting positions (expressed

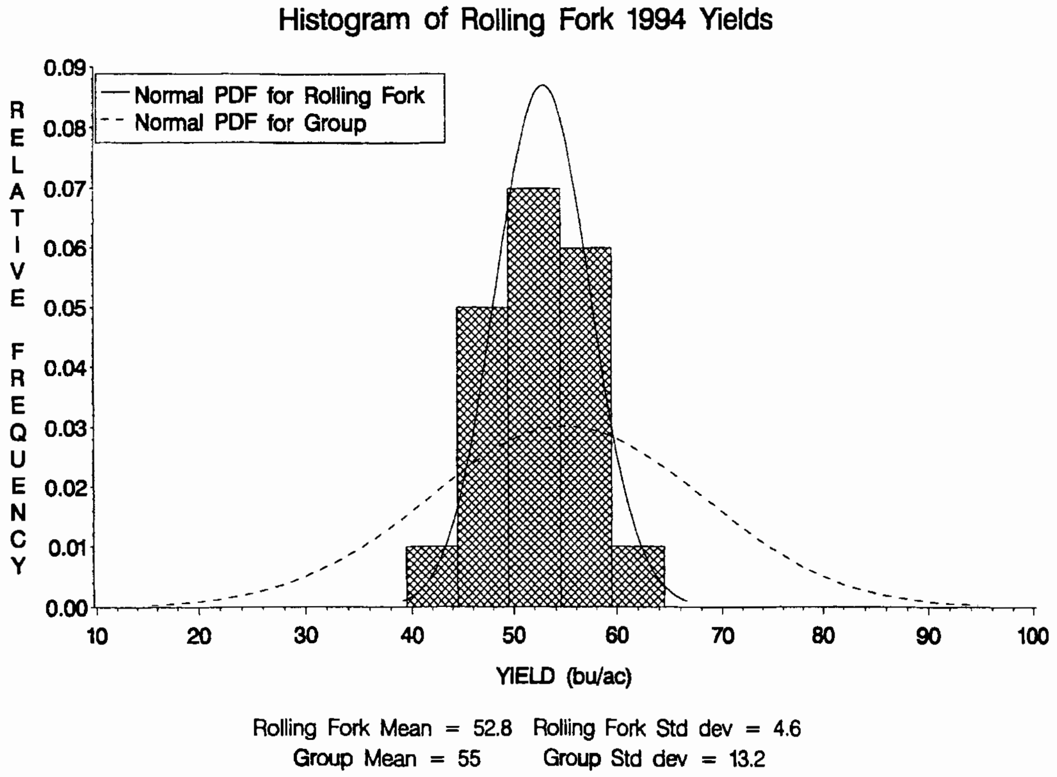


Figure 17.4 Histogram of relative frequencies of soybean yields for 20 varieties in 1994 at the Rolling Fork location. The subgroup mean and standard deviation are for only the yields obtained at this location; the group mean and standard deviation pertain to all the yields found in Table 17.1.

as a percent) using this rule for the example soybean yields are given in Table 17.4 and plotted, by hand, on Figure 17.10 using normal probability paper.

An important modification is necessary for data sets having fewer than 20 points. In these cases, the proper formula (King, 1980) for obtaining the plotting positions is the expression $x_i = (i - 0.375)/(n + 0.25)$. Small-sized samples can introduce artifacts not related to the nature of the data unless corrections are made.

Probability plotting is a graphical technique that can be used to help gain insight about a set of data. The technique can help identify if it is reasonable to assume that a set of sample data is from the same population or not. Probability plotting (1) is easy to use and versatile, (2) is best suited to data that are expensive to obtain and limited in availability or amount, (3) provides for the easy measurement of variability, and (4) helps users understand the implications of that variability in making decisions (King, 1980). The tool should be applicable to data that are obtained from situations where crop heterogeneity exists in a field, or soil fertility, water, and/or insect abundance differ across the field. Differences in the skill of field scouts that cause field-to-field differences in measured responses of interest could also be revealed (be careful here as fields could really be different). Also, these plots can communicate the results of a statistical analysis to interested parties who do not have, or do not desire, the ability to be statistically proficient but still wish to excel in management (King, 1980).

One useful feature of these plots is the ability to discover situations having **contamination** of the data. Contamination of data means that data values occur that do not represent either the original state of nature, or represents the occurrence to two different

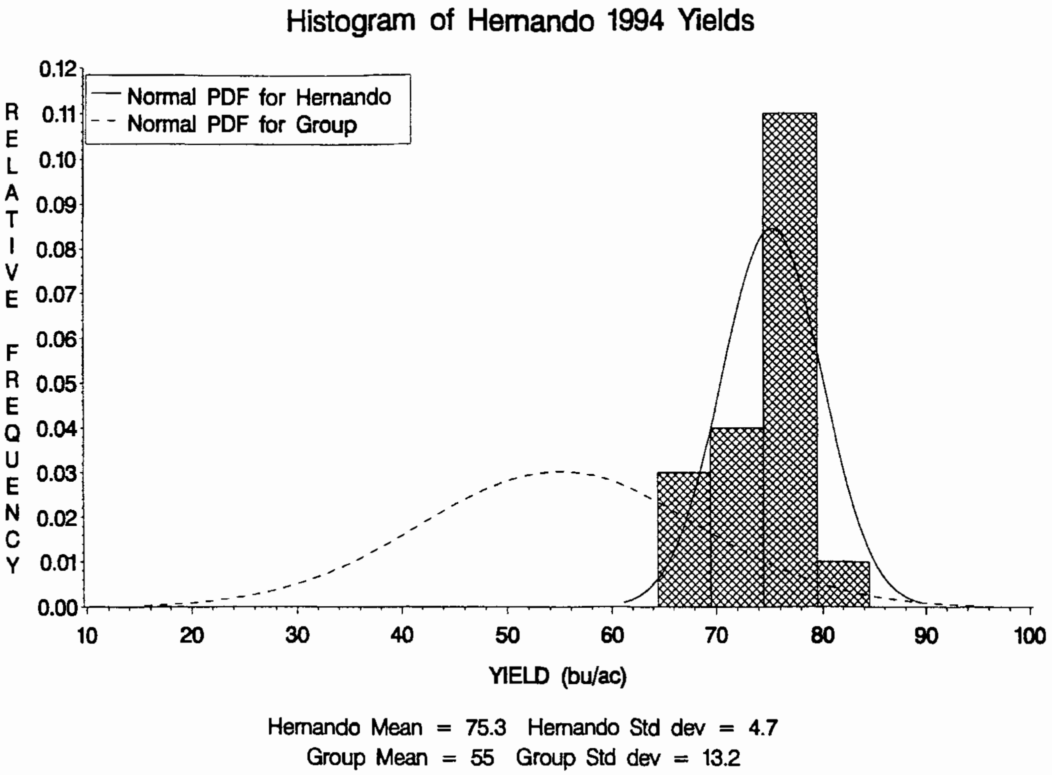


Figure 17.5 Histogram of relative frequencies of soybean yields for 20 varieties in 1994 at the Hernando location. The subgroup mean and standard deviation are for only the yields obtained at this location; the group mean and standard deviation pertain to all the yields found in Table 17.1. Yields are the highest here of all locations.

processes, situations, or events at the same time. Sometimes, contamination represents "bad" or carelessly obtained data. Whenever contamination is observed in a set of data, careful analysis must follow to determine the cause of the discrepancies. Detecting these irregularities is one advantage for drawing a probability plot for a set of sample data. When the cause of the irregularity is identified, valuable insight into system behavior has been obtained. The data of Tables 17.1 and 17.4 show in Figure 17.10 the existence of at least two populations. Yields above 60 bu/acre begin to plot along a different trend. All but one of these yields are from the Hernando location. The higher yields for the Hernando location come from another population, and therefore are governed by different "processes" than those processes that are at work at the other three locations. Earlier, the histograms of all yields (Figures 17.2 or 17.7) showed something unique about the 20 varieties at this location, but these plots could not tell if a difference was due to a different process or due merely to being larger yields. So, while a casual examination of Table 17.1 and the histograms quickly set apart the Hernando yields, it is the probability plot that conclusively shows that these yields are distinct for reasons more than just simply being larger.

The investigators who conducted the original study would have to explain probable causes. Determining these causes would be instructive. It is beyond the ability of this chapter to determine why, because not all necessary information is available. If a probability plot or histogram shows evidence of "mixed" populations in the sample for data collected, it is hoped that enough background is locally available to allow the identification of possible reasons.

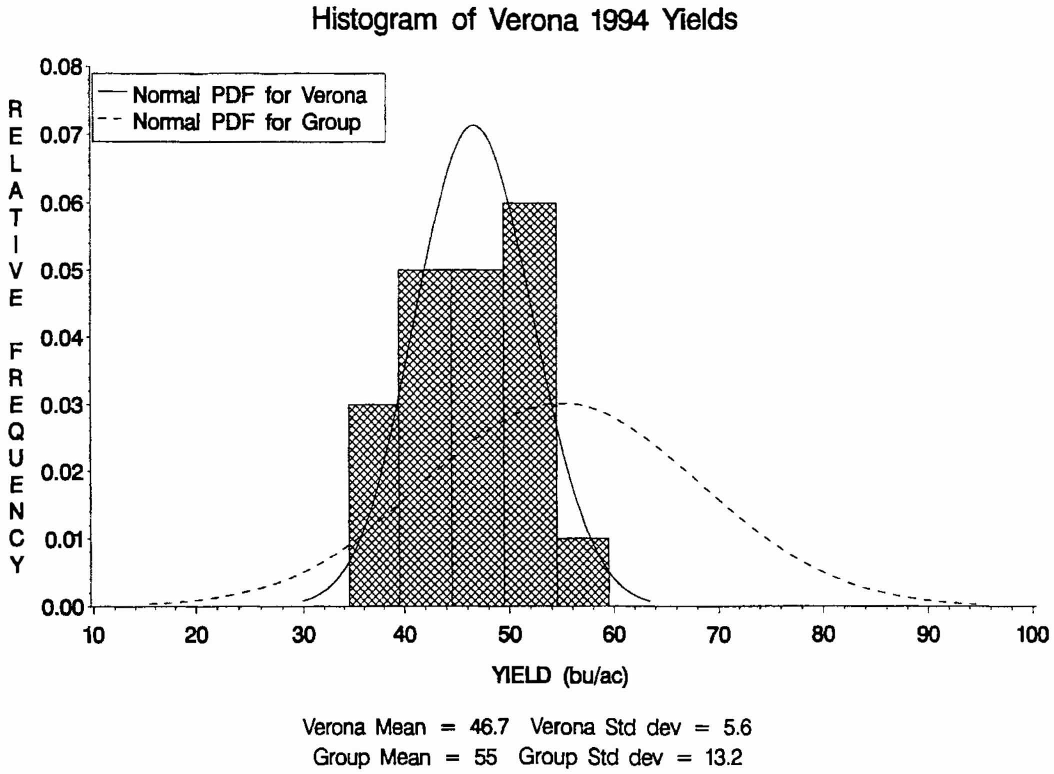


Figure 17.6 Histogram of relative frequencies of soybean yields for 20 varieties in 1994 at the Verona location. The subgroup mean and standard deviation are for only the yields obtained at this location; the group mean and standard deviation pertain to all the yields found in Table 17.1.

Interaction plots

There are two types of interaction plots. Of these, only one can be drawn without using advanced statistical techniques. The one that can be drawn without special skills, called a Type I interaction plot (Milliken and Johnson, 1989), is demonstrated here. Interaction plots are best used with data that can be arranged into a two-way table. A two-way table should be balanced, meaning that all row and column combinations (i.e., cells) have an entry. The entries in each cell should be numerical values and have like units. For example, do not include in the same table entries for some cells that are bu/acre and others plants/acre. The factors of the table can be different numerical levels (e.g., rates of nitrogen fertilizer) or different qualitative labels of the same thing (e.g., location or varieties). Interaction plots permit one to determine quickly if the observed response at different settings of one treatment interact (i.e., nonparallel responses) with different levels of a second treatment. The graphs that are shown were drawn using a computer, but the reader should note that any of these can be drawn by hand using regular graph paper, a ruler, and a pencil.

To draw a Type I plot, label and scale the y-axis to the range of the response data. Graph on the x-axis different levels (coded as indexes, 1, 2, 3, etc.) of one of the two factors. The second factor of the two-way table is used to define the different lines that are drawn on the plot. Another plot of the same kind can be drawn by reversing the order and scale of the factors on the x-axis. If too many lines on the same plot make it difficult to interpret the plot, then select a smaller set of lines and draw only those. For example, an interaction

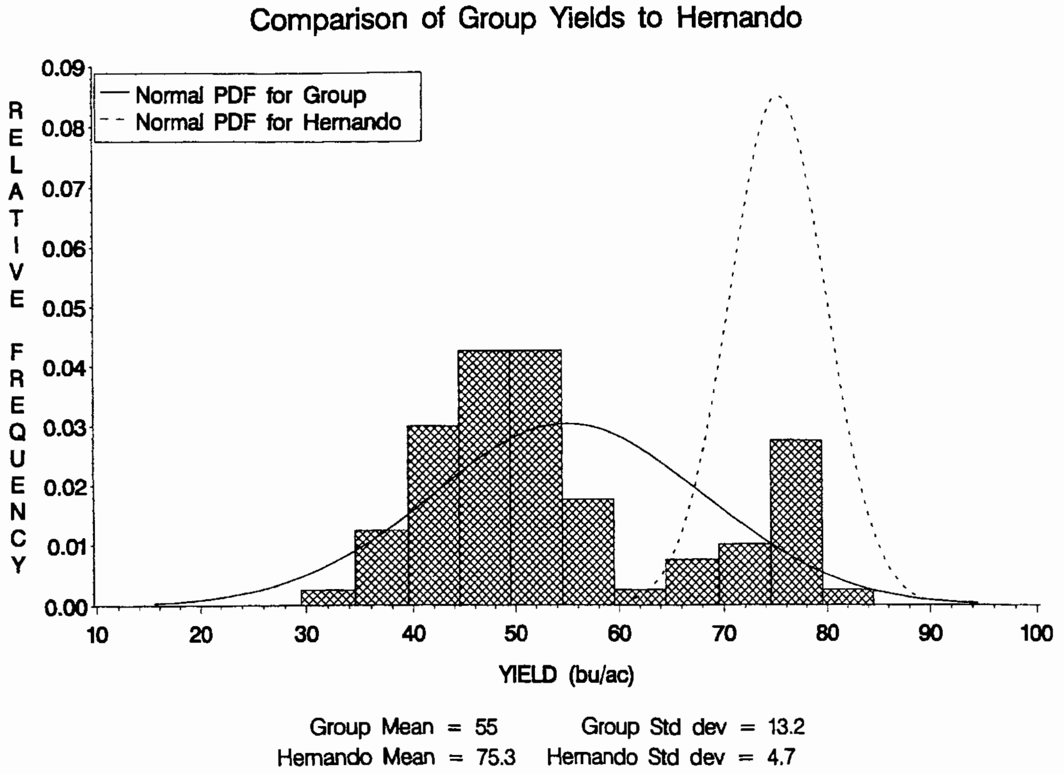


Figure 17.7 Histogram of relative frequencies of soybean yields for 20 varieties in 1994 at all locations. The group mean and standard deviation are for the yields at all locations. Shown for comparison (dotted line) is the fit of the normal density function for data from the Hernando location. Compare to Figure 17.2, which is a histogram of absolute frequencies.

plot of all 20 varieties over the four locations is too cluttered to be useful. The plots are examined for occurrences, or the lack thereof, of parallel line segments.

To keep the example plots simple, only the yields of three varieties (DP3589, H5454, and H5545) at four locations will be used (see Table 17.1). In Figure 17.11, these yields are graphed across indexes that represent the four locations. The graph indicates that the yield response of location 1 (Clarksdale) shows an increasing trend in yield. Locations 2 (Rolling Fork) and 4 (Verona) show decreasing trends, and location 3 (Hernando) shows the highest, but flattest yield response.

In Figure 17.12, another Type I plot is drawn, but here the yields across the four locations are graphed by indexes that represent the three varieties. The varieties DP3589 (line 1) and H5454 (line 2) yield similarly between Clarksdale and Hernando and differ from H5545 (line 3). But, DP3589 differs from H5454 and H5545 among the Hernando, Rolling Fork, and Verona locations.

The interaction plots very clearly illustrate the challenge of how to select the best variety for a location. A variety that performs poorly in one location may excel at another. The problem of variety selection is even more difficult if different years are considered. Advanced techniques do exist to sift the performance of varieties among locations and different years (see Gauch, 1992). These methods are not discussed in this paper.

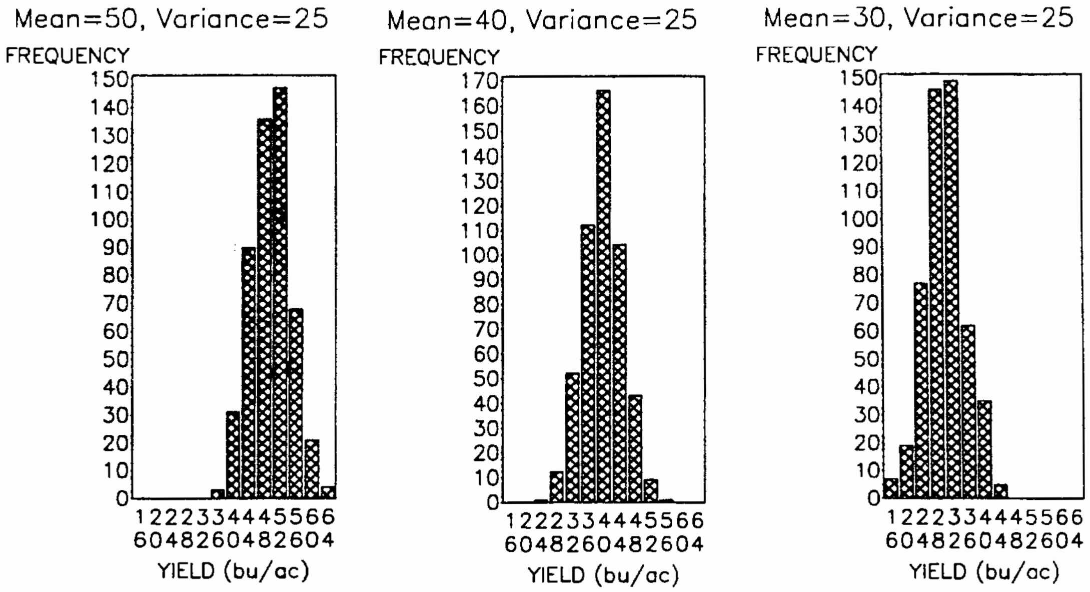


Figure 17.8 Histograms of soybean yields from three imaginary fields, each 500 acres in size, that have the same variance but different means. Reported are the yields from each acre in the field since the sum of the absolute frequencies equals 500. Notice that each histogram has a similar shape, but that as the mean increases, the center of the distribution moves to the right along the *x*-axis.

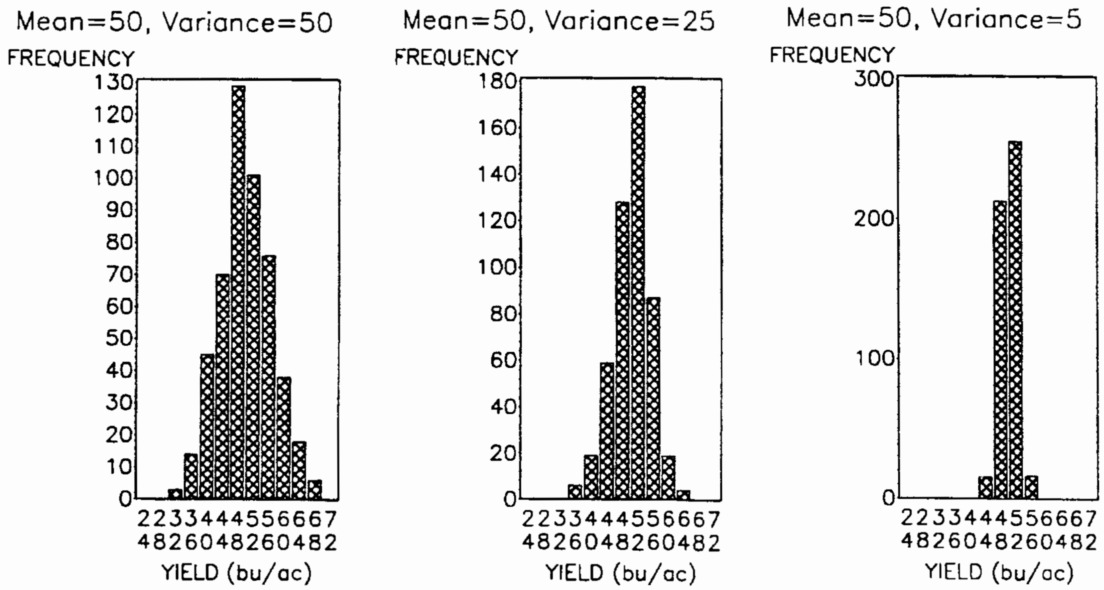
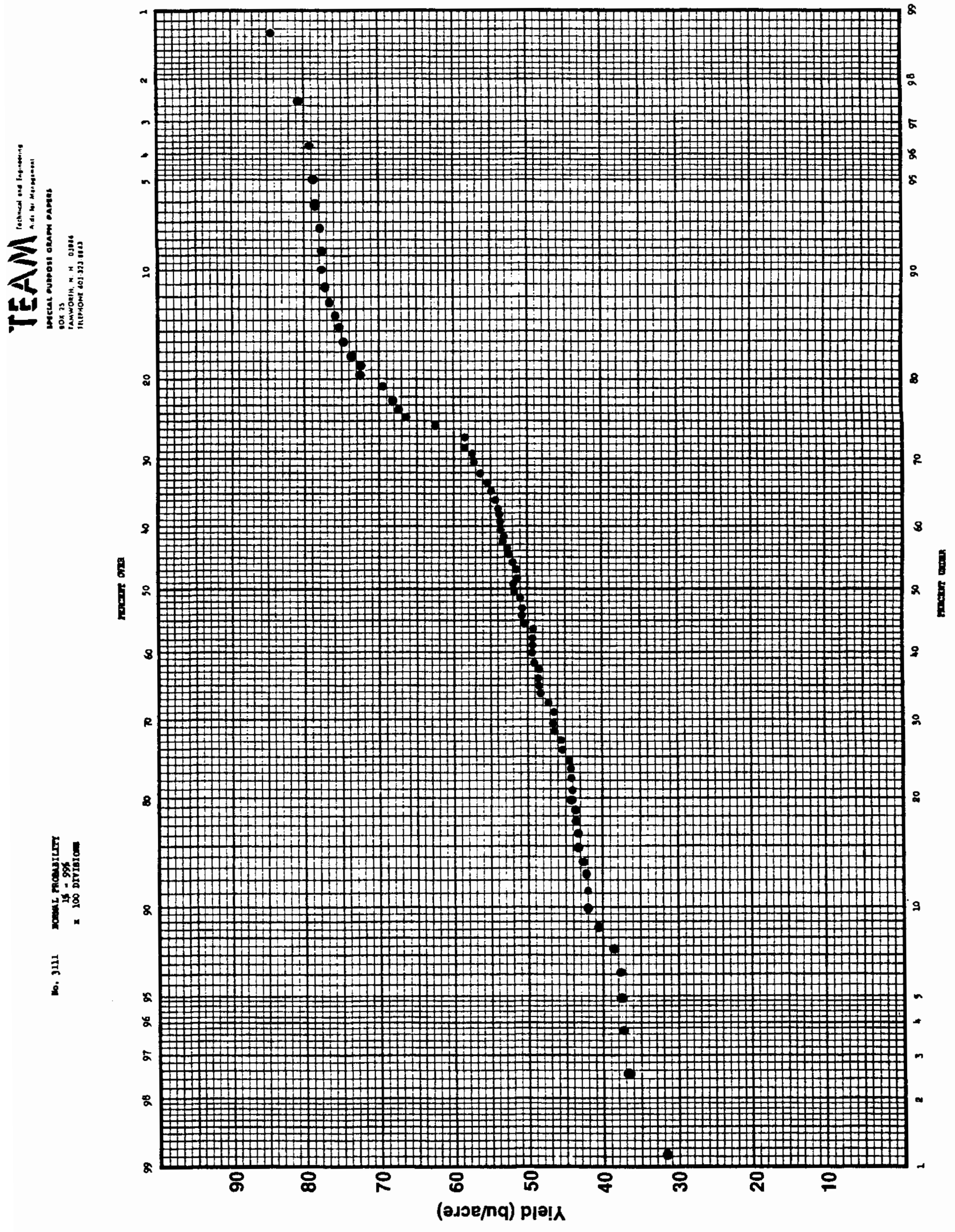


Figure 17.9 Histograms of soybean yields from three imaginary fields, each 500 acres in size, that have the same mean (or average) but different variances. Reported are the yields from each acre in the field since the sum of the absolute frequencies equals 500. Notice that each histogram has a different shape, but each distribution is centered at 50 bu/acre. Each distribution becomes broader as the variance increases.


TEAM Technical and Engineering Aids for Management
SPECIAL PURPOSE GRAPH PAPERS
BOX 25
TAMWORTH, N. H. 03886
TELEPHONE 603-323-8843
No. 3111 NORMAL PROBABILITY 1% - 99% x 100 DIVISIONS
PERCENT OVER
1
2
3
4
5
10
20
30
40
50
60
70
80
90
95
96
97
98
99
PERCENT UNDER
99
98
97
96
95
90
80
70
60
50
40
30
20
10
5
4
3
2
1
90
80
70
60
50
40
30
20
10
Yield (bu/acre)

Sampling for soil characteristics

Concepts useful to sampling soil attributes are discussed next. Again, the discussion is somewhat general and interested readers are referred to selected references for details (e.g., Wollenhaupt et al., 1997).

Effect of composite samples

It is often relatively easy to collect soil samples, but the cost of analysis of numerous samples collected on a small scale is expensive. To help keep costs down, a technique known as composite (bulk) sampling is used to estimate the average value of a soil property. In composite sampling, a fixed number of samples are collected by either a random sampling or a systematic sampling scheme. A random sampling scheme simply identifies the points on the field where soil samples are taken in a random manner. A systematic sample uses some sort of predefined grid to determine the sample points. Theoretical studies have shown that unless the value of interest follows a cyclic trend, a systematic sample can result in a more precise estimation. This has also been supported by field and laboratory studies. Once the samples have been collected, they are combined or "composited" and thoroughly mixed. The mixture is then subsampled and those samples analyzed and averaged to obtain an estimate of the average value of interest.

There are several advantages and disadvantages to composite sampling. First, the cost of analysis is typically much lower than if each sample were measured separately. If the act of creating a composite sample yields a homogeneous mixture, then the estimate is unbiased. However, if the mixing is not thorough, a biased estimate of the desired average can result. Composite sampling results in an estimate of the average only; no estimate of the variation of the property in the field can be obtained. Variability among the samples analyzed can be quantified, but this measures the thoroughness of the mixing procedure and the accuracy of the test procedures rather than the variation in the field itself. Since values in the field can vary substantially even for samples taken near each other, reliance on the estimated average value without consideration of the variation in the individual samples can lead to overestimation of the quantity of interest in some areas of the field and underestimation in others. When these estimates are used to provide guidelines for nutrient requirements, significant over- or underapplication can result. Research has also suggested that the act of compositing the samples can change the physical composition of the soil in such a way as to alter the values of some soil properties (Giesler and Lundström, 1993).

Historical overview of soil sampling and interpolation methods

Soil testing has been used since the late 1940s to identify soils that may require lime and fertilizer inputs for optimizing crop production (Bray, 1929; Truog, 1930; Morgan, 1932). For plant nutrients, it involves rapid chemical analyses in addition to interpretation, evaluation, and fertilizer recommendations (Peck and Soltanpour, 1990). With the increasing awareness of fertilizer effects on environmental and soil quality, soil tests can also be used to determine where fertilizers or manure should not be applied as well as where to apply them.

Figure 17.10 Example graph of normal probability plotting using special-purpose graph paper. The data are the yields of Table 17.1 after being arranged in rank order (see Table 17.4). The plot is a reproduction of one drawn by hand.

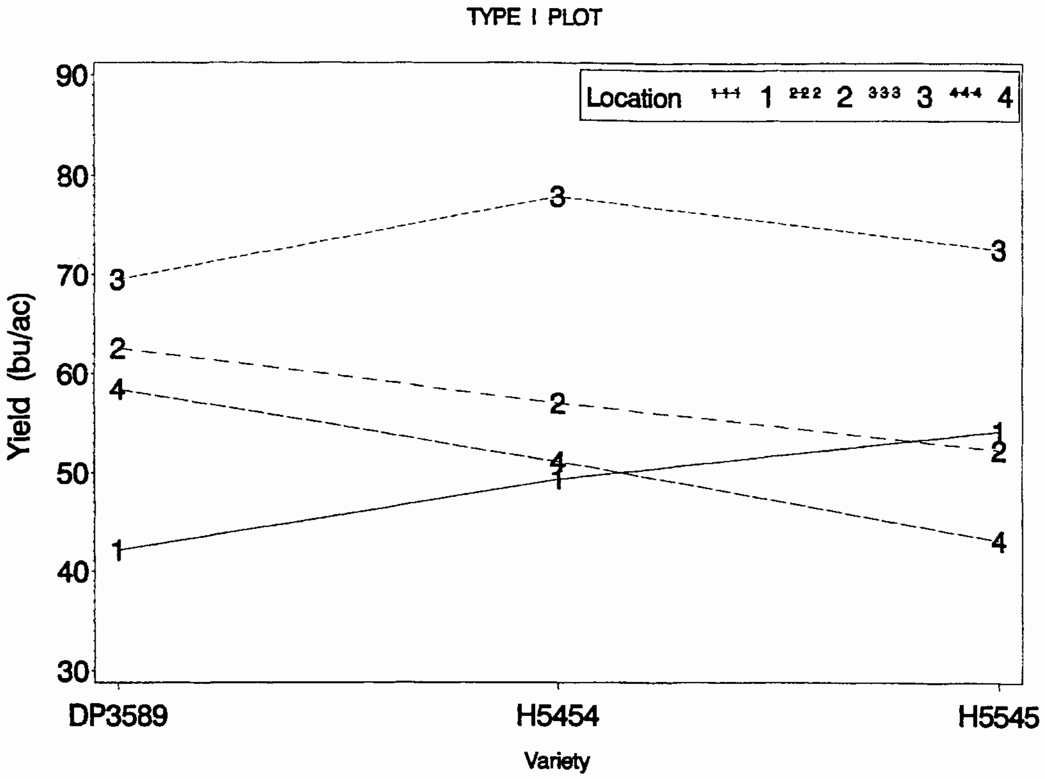


Figure 17.11 Type I interaction plot for the yield response of three varieties (see text) at each of four locations. Parallel line segments indicate where the yield response of a variety is similar to another between locations. Line segments that are not parallel indicate an interaction in yield response of the varieties over locations. Line 1 is Clarksdale, line 2 is Rolling Fork, line 3 is Hernando, and line 4 is Verona.

Soil testing has historically focused on determining the average soil test value for a field or area. It assumed that each observation was independent from other observations, and based on that assumption many chapters and articles have appeared in the literature. Cline (1944) presented general principles of soil sampling that were expanded by Peterson and Calvin (1982) and James and Wells (1990).

Most soil sampling efforts focused on determining an adequate number of samples to provide a reliable estimate of the mean, the most efficient sampling plan, and some measure of spatial variability. Peterson and Calvin (1965) defined the best sampling plan as one that gave the lowest sampling error at a given cost or the lowest cost at a given sampling error. However, past sampling research has shown that grid sampling almost always increases precision compared with random sampling due to the spatial correlation of values (Peck and Melsted, 1967; Sabbe and Marx, 1987). Most producers and agribusinesses have done composite (or bulk) sampling to determine field averages.

The underlying probability distribution functions of many soil parameters are usually not normally distributed but are log normal (Reuss et al., 1977; Parkin et al., 1988; Hergert et al., 1997). If data are distributed log normally, more than 50% of the values are less than the mean (Figure 17.13). A few high testing values can skew the mean and cause an overestimation of the central tendency. This fact probably has led to more confusion and questioning of soil test credibility than any other factor. This was not a major consideration in past soil test correlation/calibration research because plot sizes were generally small

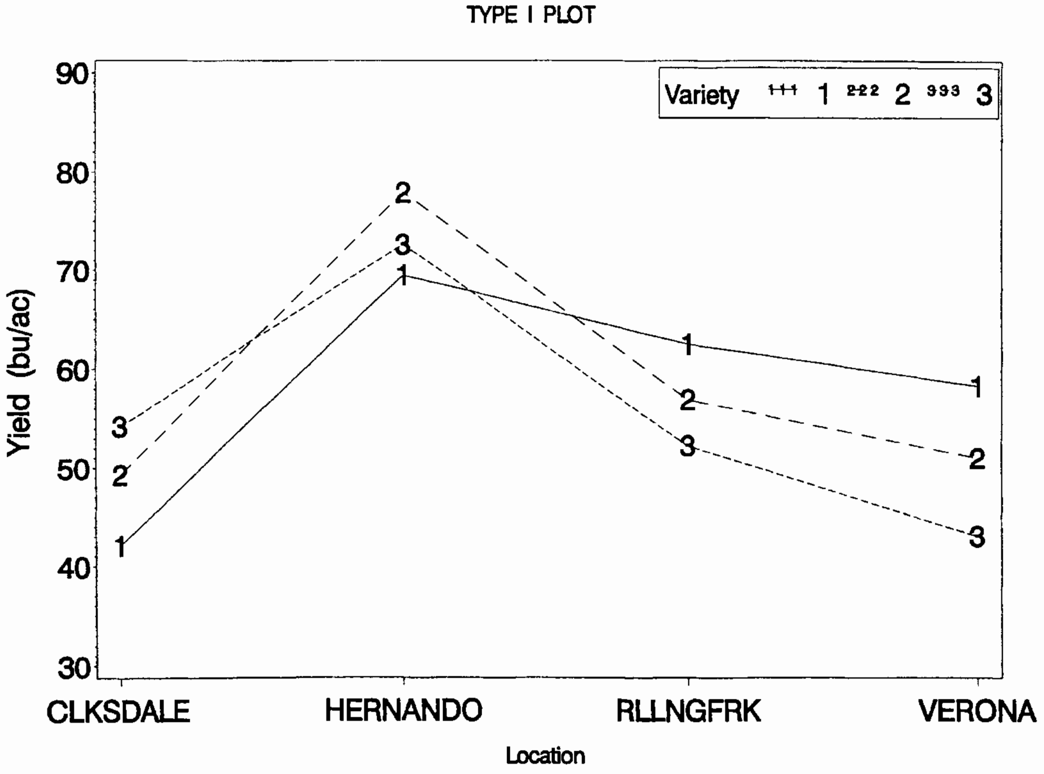


Figure 17.12 Type I interaction plot for the yield response by variety at four locations. Parallel line segments indicate where the yield response of a location is similar to another location for the selected varieties. Line segments that are not parallel indicate an interaction in yield response for different locations. Line 1 is the variety DP3581, line 2 is H5454, and line 3 is H5545.

and homogeneous so the critical levels that were developed were not the problem. The problem was one of scale when small-plot information was extended to the field scale, which encompassed much wider variability (Hergert et al., 1997).

Advances in the theory of regionalized variables (geostatistics) enables estimation of the spatial dependence of soil properties regardless of the underlying distribution (Matheron, 1971). Geostatistical analysis provides a method to develop "statistically correct" contour maps of the soil parameter being measured regardless of the underlying frequency distribution. Many additional samples are required, however, compared with classical sampling methods. The advent of site-specific management (SSM) and variable-rate application (VRA) has caused many producers to change their soil sampling methods.

The dilemma is that quantifying the variability of a soil test parameter requires soil sampling at an intensity that will allow the variability to be mapped spatially with some degree of confidence. This is nothing new, as Reed and Rigney (1947) concluded that field variation was much greater than laboratory variation. Each soil property has a unique variation in a specific field, and the specific soil property having the greatest variation could not be anticipated. Soil properties show large differences in spatial and temporal variability. In general, soil properties with a larger variability require more-intensive sampling to quantify the pattern of variability. Wollenhaupt et al., (1997) lists common differences between soil test parameters and their spatial range.

Sample properties can also be classified according to temporal variability. Soil properties that do not change appreciably over many years include organic matter content,

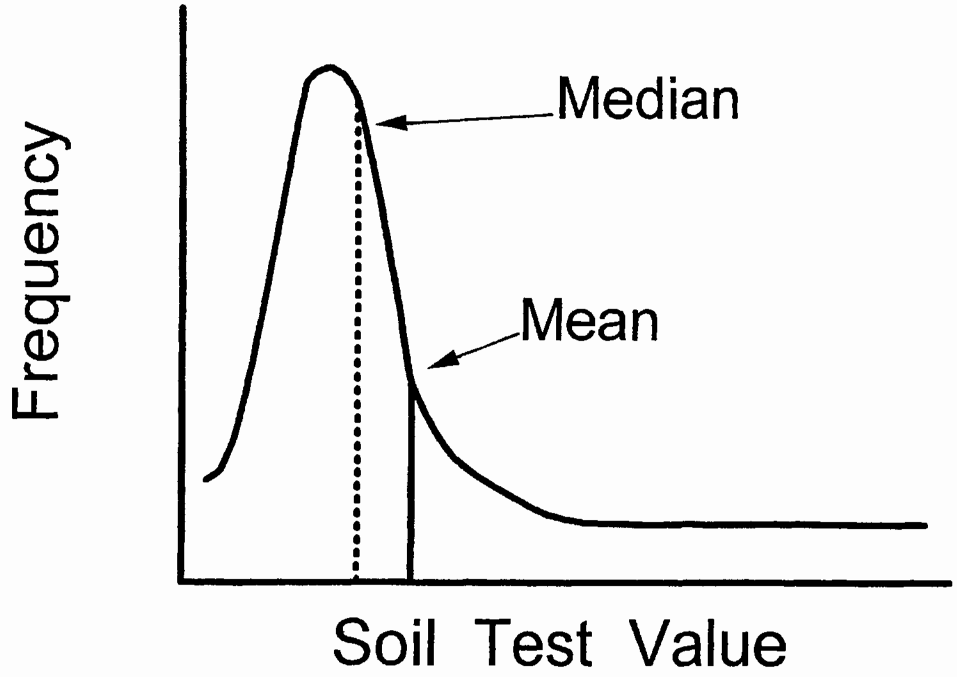


Figure 17.13 Diagram showing the general distribution of many soil characteristics. Notice that the curve is not a normal distribution since it is not symmetrical about the mean and that these data are skewed to the right.

texture, and cation exchange capacity. Soil pH changes slowly over periods of 5 to 10 years. Soil P and K may show dramatic changes over 3 to 5 years, whereas soil nitrate or sulfate can show large changes within one growing season or year. Sampling time and frequency need to be guided by the dynamics of the soil property being measured. Proper timing of sampling is critical for accuracy when measuring dynamic vs. static parameters.

Soil and fertilizer management for crop production has increased field variation in soil nutrients and pH. Most past soil sampling guidelines called for sampling the depth of tillage, normally 6 to 8 in. (Peck and Melsted, 1967). However, fertilizer application methods (band vs. broadcast, manure application, differential crop removal) and changing tillage practices have increased heterogeneity and complicated both the sampling and interpretation process (Randall, 1982; James and Wells, 1990; Kitchen et al., 1990). Additional research will be required to determine future suitable sampling techniques to deal with nutrient and pH stratification.

Fertilization according to VRA and SSM provides additional challenge and opportunity for improved soil sampling techniques. The assumption of a VRA is that it will more closely match productivity, input efficiency, and profitability compared with uniform application (Sawyer, 1994). The concept assumes that within-field variability exists, that it influences crop yields, that the variation can be identified, measured, and mapped, that precise crop-response models are available to determine optimum economic inputs, and that data processing and application equipment are available that can effectively manage and apply the inputs.

This new methodology suggests reevaluating past research and assumptions about soil sampling. We know from past research that uniform fertilizer management across an entire field can result in overfertilization of some areas and underfertilization of others. In either case, this is an economic loss to the producer. Overfertilization increases the probability of nutrient loss from erosion or leaching and may accentuate environmental problems, whereas underfertilization limits yield and product quality.

The focus of this section is a discussion of soil sampling techniques from two perspectives — classical sampling to determine average soil properties for use in whole field management (WFM), and sampling and interpolation techniques that can be used to develop precise maps for SSM.

Sampling frequency

Sampling frequency should be based on expected changes in the soil parameter as discussed earlier. For many properties or parameters, periods of 4 to 5 years will be adequate, whereas mobile nutrients like NO_3–N and SO_4–S may require yearly sampling to detect changes. This recommendation would apply whether sampling is based on WFM or SSM. The challenge in both systems will be to determine the impact and interaction of other factors including uniformity or lack of it, sampling pattern, number of samples, and sampling depth as influenced by past or current tillage.

Sampling methods

The degree of variability for the different scales the producer plans to manage should guide sampling. Soil variability can be classified into three categories of variability, ranging from the smallest to largest: micro-, meso- and macrovariability (see James and Wells, 1990). With WFM, producers are managing above the macroscale by incorporating micro-, meso-, and macrovariability into a bulk (composite) sample that represents the field average. This may not be a major limitation if the field is fairly uniform. SSM is more concerned with variability on the meso- and macroscales, and is most helpful where macrovariation is large.

Sampling uniform fields

Uniform fields have small macro- and mesovariability (James and Wells, 1990). Similarities in slope, elevation changes and drainage, aspect and management history including fertilizer and lime application, and cropping would be considerations to determine if a field may be categorized as "uniform." For these types of fields, WFM and classical random composite soil sampling is still a reliable tool. The value of information gained from extensive and expensive grid soil sampling, the possible savings on fertilizer or lime, and increases in productivity probably would not offset the additional time and money required for a more-intensive sampling (Franzen and Peck, 1995a; Gotway et al., 1997).

Sampling guidelines recommending areas no larger than 40 to 50 acres per sample with a minimum of 25 to 30 cores per area apply for these types of fields for most nutrients and soil properties (Sabbe and Marx, 1987; James and Wells, 1990). A simple random sampling pattern is favored by most agronomists, although the literature suggests other systems including zigzag patterns (Sabbe and Marx, 1987). The unknown is the influence of tillage on required depth of sampling. Changing tillage and fertilizer application techniques have complicated sampling by stratifying nutrients and pH. Variability with depth exists but what has not been determined is how plants respond to these highly variable conditions. Do they simply adapt to the "average" that would be expressed by 0 to 8 or 12 in. samples or is there an effect of stratification on the crop?

Current guidelines suggest shallower sampling depths of 2 to 4 in. can replace the traditional sampling depth of 8 to 12 in. (James and Wells, 1990). This applies for no-till and ridge-till fields for monitoring surface pH and the buildup of immobile nutrients.

Sampling for site-specific management

There is some uncertainty about the exact sampling frequency, method, and pattern required to develop maps for SSM that show significant variation at a reasonable cost compared with classic bulk soil sampling. Information comes at a cost. Producers cannot expect to spend the same amount of money for SSM as for WFM. However, the value of the information should offset the inc:eased soil sampling cost. Because of the intensity that will be required for SSM soil sampling, farmers must first look at sampling frequency. For many properties, one sample every 4 to 5 years will be sufficient. This allows yearly sampling of 20 to 25% of the land area per year.

Another factor that has changed soil sampling is who does the sampling. When operations were smaller, many farmers were thoroughly aware of their fields. They have a distinct advantage in that they have prior knowledge about productivity, problem areas, etc. that can help them direct some of their sampling. As farms have grown larger, much of the soil sampling has been done by agricultural consultants and fertilizer/agricultural chemical company personnel. These individuals may not always have this prior knowledge. Knowing the production history of a field can help guide sampling. This is part of the reason for the excitement of generating yield maps. The exact relationships of productivity as a basis for soil sampling pattern, however, still needs to be established by research. Sharing this information between the producer and an agricultural consultant or a fertilizer dealer would be helpful in developing sampling plans.

A number of recent papers have looked at sampling intensity (Hergert et al., 1995a,b; Franzen and Peck, 1995b; Wollenhaupt et al., 1997). Although there is no single recommended intensity, there is some agreement that sampling units at spacings above 196 to 230 ft causes a loss of information. Simply stated, the more points used to make a map, the better the map. If there is limited background information about a field, some type of systematic sampling is suggested. There are a number of different options, including aligned or unaligned grids. Variations include taking one large core from the square that represents the area, or possibly taking four to eight cores in a random pattern within the square.

Another factor to consider in the sampling pattern is the interpolation technique that will be used to produce a map from the data. Numerous interpolation techniques are available (Wollenhaupt et al., 1997). Most interpolation techniques accommodate data that are not equally spaced. This may be an advantage for some soil properties where there is some idea of the variability. For example, soil organic matter can be mapped based on a bare soil from using aerial photography. Intensive sampling of a corresponding location (Gotway et al., 1997) developed an excellent map with close correspondence to soil organic matter. However, using the bare soil photograph as a guide, a much smaller directed sampling could have been used to determine the same information. In the case of soil organic matter, taking soil samples close to boundaries provides additional information that when interpolated describes changes between soil zones.

For many other soil properties (P, K, pH), there is usually little background information about where to sample. In these situations the best sampling plan is to use some type of regular grid making sure that there are sufficient points to develop a good map. As mentioned previously, the sampling plan can influence the interpolation technique. Most interpolation techniques can handle non-uniformity spaced data.

As more sampling is done for SSM, the selection of which interpolation techniques to use becomes a factor. The selection of the interpolation technique probably will be much less important than taking a sufficient number of sample locations to produce a good map. Sampling intensities that require at least one sample point for each 1.25 acre would be ideal, although samples for areas up to 2.5 acre may be adequate. There is considerable debate concerning these issues, but the bottom line is that information does come at a cost. Therefore, to do a better job of managing a crop, it must be remembered that if you can not measure it, you can not manage it.

Line-intercept sampling for stand analysis

The LIS method for obtaining information about stands was adapted from long-established techniques used in forestry and wildlife biology (Kaiser, 1983). This method has been little used in row crop agriculture, but recently its use has been advocated for stand analysis and scouting for insect pests in early-season cotton (Willers et al., 1992; Williams et al., 1995). A closely related application has been the estimation of crop residue for erosion control. Samples are drawn from row segments of equal length on each row crossed by the transect line (Williams et al., 1995). If seed are broadcast, the method cannot be used as described here, because there will be no parallel rows. (For broadcast fields, a technique can be developed. If this is a need for any fields on your farm, contact the lead author, JLW, and a protocol will be developed). In this discussion, the method estimates only one attribute — the number of soybean plants per acre. Other attributes can be estimated, but the technique has not been modified for other uses in soybean production.

The first step in using LIS to estimate the stand of a soybean field is to divide the field into subunits (strata) where crop phenology is similar by applying concepts analogous to the use of stratified sampling plans. Here, these smaller divisions of a field are called management units. Depending upon the situation, a good range in size for a management unit is between 50 and 100 acre. It is recommended that at least one sample line be used per management unit and, if time is available, up to four lines per management unit. Generally, each line should be at least as long as the width of one planter pass and no longer than the width of four planter passes. As the row spacing of the crop gets narrower, it becomes more convenient to use only one line per pass. The other remaining requirement is that the sample lines should not overlap.

A starting point for a single sample line is chosen at random in the management unit. In soybeans, it is best if the starting point be midway between the first or last drill of one pass of the drilling machine (seeder or planter) and the first or last drill of the next adjacent pass. Sampling is conducted along the transect line across consecutive rows. The total number of plants in a fixed length of row is counted on each row crossed by the transect line. Typically, data are collected from 1- to 3-ft sections of row. If the number of plants per foot is large and there are not too many gaps in the drill, the 1-ft length is sufficient. If the number of plants per row is variable, then 3 ft should be used. If the stand is extremely sparse, one could even use a 5-ft sample length per row. The process of moving across rows gives this method its strength by capturing the variability in the crop due to planting irregularities that occur among planter boxes and other causes that vary the number of plants in each row.

The estimate of plants per acre depends on the variability of the stand in each row crossed by the transect line, the length of row sampled, and number of transect lines used per field. Do not change the length of row sampled on each row for any lines in a management unit; that is, do not sometimes use 1 ft of row and another time use 3 ft and another time use a 5-ft sample on the same line. Also, make sure the transect line is as straight as possible. Do not let the line become crooked or vary too far from being perpendicular to the row direction. Some people actually stretch a small cord or rope between two stakes and obtain a very straight line. Then, they lay one end of the yardstick or ruler against the rope on each row and make their count.

Table 17.5 LISs for Estimating the Number of Soybean Plants per Acre

Row	Sample variates	Sorted variates	Cumulative sum	Cumulative %
1	6	1	1	0.005
2	8	3	4	0.022
3	12	3	7	0.038
4	10	3	10	0.054
5	1	4	14	0.076
6	9	5	19	0.103
7	10	5	24	0.130
8	5	6	30	0.163
9	6	6	36	0.196
10	11	8	44	0.239
11	3	8	52	0.283
12	8	8	60	0.326
13	10	9	69	0.375
14	10	9	78	0.424
15	5	9	87	0.473
16	9	10	97	0.527
17	4	10	107	0.582
18	12	10	117	0.636
19	10	10	127	0.690
20	3	10	137	0.745
21	12	11	148	0.804
22	9	12	160	0.870
23	3	12	172	0.935
24	8	12	184	1.000

Average/3 ft = 7.67	Standard deviation/3 ft = 3.25	Range/3 ft = 1-12 plants
Average/1 ft = 2.56	Number row-feet/acre = 37,336	Estimated plants/acre = 95,580

Reported are the number of plants per row for each of 24 drills in one pass of the seeder, at 14-in. row widths. The fixed row length per sample on each drill was 3 ft. In practice, at least four drill passes per managment unit should be sampled. See text for further explanation and compare with Figure 17.14, which graphs the empirical distribution function (ECDF) of these data.

Calculations

If the rows are parallel, the only number needed is the total number of linear row-feet per acre for that row spacing. This value must be known in order to calculate a per-acre estimate of the number of soybean plants using the line-intercept method. Mississippi Cooperative Extension Service Bulletin Publication 883 reports the number of linear row per acre for several different row widths from 6 to 40 in.

Presented in Table 17.5 is an example data set of a 3-ft sample across 24 rows of a single pass of the seeder from a field of soybeans in the Mississippi Delta during 1996. The row spacing was 14 in., which means, in round figures, there are 37,336 linear row-feet per acre. In this 24-row sample, a total of 184 soybean plants were counted for this transect line. The total number of row-feet examined in the sample is 72 (24 rows × 3 ft/row). Dividing 184 plants by the total number of row-feet sampled gives an estimate of 2.56 plants per foot of row. This value when multiplied by the number of linear row-feet per acre for this row spacing provides an estimate of 95,580 plants per acre.

Table 17.5 also reports the final result and presents several descriptive statistics based upon the 3-ft sample units along the transect line. The table also shows the sample sorted

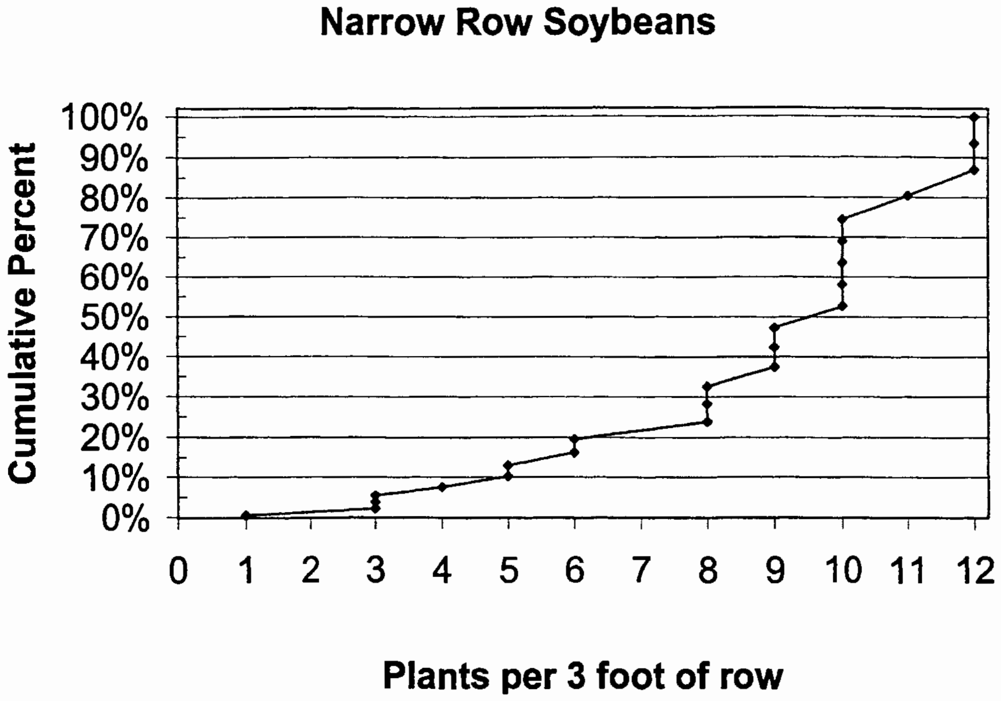


Figure 17.14 Diagrammatic distribution of stand variability for narrow-drill soybeans. Counts were obtained from one 24-row-long transect line where the number of plants was counted in each 3-ft length for each row. If the stand where uniform, the plot would be a single vertical line at one of the values for plants per 3 ft of row.

into ascending order, the cumulative sum of the values, and the **cumulative percentage**. The cumulative percentage is obtained by dividing the cumulative sum column by the total. The cumulative percentages can be graphed against the sorted observations (Figure 17.14). This graph represents the within-row variability in stand for the field. The median number of plants per 3 ft of row is the value on the x-axis where the cumulative percentage is 50. For these data (Table 17.5) the estimated median is 9.5 plants. The mode is 10 plants per 3 ft of row which can be seen easily in Figure 17.14 since this value has the most points. Note from Table 17.5 that the estimated average is 7.67 plants per 3 ft of row. When the mean is smaller than the median and the mode, the distribution of the sampled attribute is skewed left.

If several lines are collected in a management unit, the average for that unit is the average of the averages estimated for each line. For example, if four transect lines were sampled in one management unit and provided estimates (rounded to the nearest whole plant) of 88,165, 86,958, 91,355, and 89,753, the average using Equation 17.1 would be 89,058 soybean plants per acre. The estimated standard deviation after taking the square root of the result of Equation 17.2 is 1912. By the use of the information in Table 17.2, the estimate of the standard deviation is (91,355 – 86,958)/2.059 = 2135. Again, as before, the short-cut approach agrees favorably with the results given by the equation. Armed with these results, the farm manager can make appropriate management policy for any decision that involves knowledge of the stand. The method described here to obtain the stand estimate of soybeans is not difficult to employ. It is a simple process and does not require tremendous amounts of time to collect the data from one sample line whose length is the same as the width of one planter pass.

Compendium of soybean sampling literature

Several reference texts describing entomological sampling methods are available for soybeans. Two major reference books by Kogan and Herzog (1980) and Pedigo and Buntin (1993) are extremely thorough. McDonald and Manly (1989) have also written a technical text on sampling. A recent handbook edited by Higley and Boethel (1994) has appeared that includes a black-and-white pictorial key to insects and the damage they cause at different stages of soybean growth. Numerous color plates showing the adult and most immature life stages of soybean insects (including beneficial species) are displayed. The handbook also provides information on the life cycle and biology of soybean pests including range maps showing the distribution of the species in North America.

Sequential sampling plans for several insects that attack soybeans have been described. These plans include ones for green cloverworm (Hammond and Pedigo, 1976), velvetbean caterpillar (Strayer et al., 1977), and other defoliators (Bellinger and Dively, 1978). Waddill et al., (1974) describes a sequential plan for beneficial insects, including several species of the genera *Nabis* and *Geocoris*. Additional details about sequential sampling can be found in several chapters of Pedigo and Buntin (1993) and Kogan and Herzog (1980). Two bibliographies of sequential sampling plans for insects on several crops, including soybeans, are Fowler and Lynch (1987) and Pieters (1978).

From time to time, research groups put together bulletins on specialized topics. Longstanding examples are the various bulletins in the Southern Cooperative Series that are edited and authored by researchers affiliated with the agricultural experiments stations of different states. One handbook discussing the sampling of Heliothine moth pests on soybeans (along with cotton, corn, and other crops) is Southern Cooperative Series Bulletin No. 231 (1979). Another, Southern Cooperative Series Bulletin No. 377 (1994), describes soil sampling procedures for the southern region of the U.S. Peck and Melsted (1973) and Wollenhaupt et al., (1997) are other useful references on soil sampling.

Other publications such as state extension service information sheets and publications on numerous sampling issues are also available. The title and number of these small publications produced by the Mississippi Cooperative Extension Service (along with a few from other states) are listed at the end of the reference section.

Many sampling plans described from a general biology or wildlife point of view have been described in an excellent text by Thompson (1992). Discussed are basic concepts and simple random, stratified, cluster, survey, line-transect, and capture–recapture sampling plans. Spatial sampling plans (including a technique known as kriging) are also discussed. Also, several commercial companies are making information sheets that contain useful information. For example, The Potash & Phosphate Institute (Suite 110, 655 Engineering Drive, Norcross, GA 30092-2837; Telephone: 770-447-0335) has a small flyer entitled "Site-Specific Nutrient Management Systems for the 1990s" (Item #01-1180) that presents an overview and generalized recommendations for soil sampling for precision agriculture.

We have pointed out several times how computers are bringing dramatic changes in data collection, manipulation, and analysis. Similar changes are at work in the area of sampling. One method of sampling, as well as an analysis technique, is called the "Bootstrap." It is briefly discussed by Efron and Tibshirani (1991). This method, also known as resampling statistics, is popularized in a text by Simon and Bruce (1993). Resampling methods may prove to be extremely useful to those who do not have specialized training in statistics.

Other innovative sampling techniques are computer intensive (Schmitt, 1969; Plant and Wilson, 1985; Nyrop et al., 1986; Binns and Bostanian, 1990; Willers et al., 1990a,b;

Binns et al., 1996). Recently, adaptive sampling techniques applicable to rare populations have been described (Thompson, 1992). The chief advantage of many of these methods is that they utilize smaller-sized samples, or are more efficient at allocating sampling effort than many traditional methods but at the cost of a greater computational burden. Traditional methods use larger sample sizes, but often can be analyzed with a pencil and paper, or pocket calculator, hence, their greater popularity for the present time.

We anticipate that as speech recognition (Schindler, 1996), sensor development, computer advances, and other technologies continue to decrease in cost (McKinion, 1992), sampling techniques that are accurate, timely, and cost-efficient with small sample sizes will be developed. The trend toward the invention, adoption, and application of innovative sampling techniques will be pushed by increasing labor and chemical control costs.

Concluding remarks

We have described several concepts that require the use of little more than a pencil, ruler, graph paper, and handheld calculator. This information is what we hope you will begin to use. Therefore, you are encouraged to use some of these methods with data obtained from fields you manage.

Without a doubt, computers are going to be used more and more in soybean production in the years ahead. The development of better methods of retaining farm production records should also follow. Historical records that are complete, well documented, and retrievable are also sources that can be "sampled or mined" for information to produce better crops. However, to use best most material discussing the analysis of sample data, one should seek additional training and expert advice, and invest in more-sophisticated software packages. In the future, we anticipate that many techniques will be managed by "expert systems" that place the technical burden on the software and not the user. Users of such software (often called a decision support system) can instead focus their efforts on what is important — the production of an excellent crop at the lowest cost and in an environmentally acceptable manner.

There are three major points that we hope you never will forget. First, when embarking on a sampling effort, have a well-defined goal or question in mind. Understand the advantages and limitations of the sampling design used to answer your question. If you do not perform the sampling yourself, ask the person who collects the sample to tell you these things. If they do not know, ask that they find out and then let you know. The costs of production today are too high and profit margins too thin to let shoddy sample data be used to make management decisions. With respect to the sampling goal, establish well-defined sampling units and sample sizes and clearly know what attribute of the population is being sampled. In a sense, wed the biology of the attribute of interest to the appropriate sample plan. Be clear in your mind what population is being sampled. Second, be aware of variability. Do not combine samples carelessly, especially if the sampled attribute is being measured on distinct populations. Do not take an average of different attributes unless you can interpret what that average means. Third, learn and remember the distinction between the mode, median, and mean or average. If the sampled data are heavily skewed, which one of these three measures do you wish to use and which one should you use? By being wise in using and investing your sample resources, perhaps you may discover some hidden treasure in your own data that will help you better manage your soybean crop.

References

Bellinger, R. G. and G. P. Dively. 1978. Development of sequential sampling plans for insect defoliation on soybeans, *J. N.Y. Entomol. Soc.* 86:278–279.

Binns, M. R. and N. J. Bostanian. 1990. Robust binomial decision rules for integrated pest management based on the negative binomial distribution, *Am. Entomol.* 36:50–54.

Binns, M. R., J. P. Nyrop, and W. Van der Werf. 1996. Monitoring pest abundance by cascading density classification, *Am. Entomol.* 113–121.

Bray, R. H. 1929. A Field Test for Available Phosphorus in Soils. Ill, Agric. Exp. Stn. Bull. 337:589–602.

Buntin, G. D. 1993. Developing a primary sampling program, in *Handbook of Sampling Methods for Arthropods in Agriculture,* Pedigo, L. P. and G. D. Buntin, Eds., CRC Press, Ann Arbor, MI, 99–115.

Calvert, W. S. and J. M. Ma. 1996. *Concepts and Case Studies in Data Management,* SAS Institute, Inc., Cary, NC.

Cline, M. G. 1944. Principles of soil sampling, *Soil Sci.* 58:275–288.

Cochran, W. G. 1956. Design and analysis of sampling, in *Statistical Methods,* 5th ed., G. W. Snedcor, Ed., Iowa State College Press, Ames, 489–523.

D'Agostino, R. B. and M. A. Stephens. 1986. *Goodness-of-Fit Techniques,* Marcel Dekker, New York.

Efron, B. and R. Tibshirani. 1991. Statistical data analysis in the computer age, *Science* 253:390–395.

Fogiel, M., Ed., 1985. *The Statistics Problem Solver,* Research and Education Association, New York.

Fortner, B. 1995. *The Data Handbook. A Guide to Understanding the Organization and Visualization of Technical Data,* 2nd ed., Springer-Verlag, New York.

Fowler, G. W. and A. M. Lynch. 1987. Bibliography of sequential sampling plans in insect pest management based on Wald's sequential probability ratio test, *Great Lakes Entomol.* 20(3):165–171.

Franzen, D. W. and T. R. Peck. 1995a. Field soil sampling density for variable rate fertilization, *J. Prod. Agric.* 8:568–574.

Franzen, D. W. and T. R. Peck. 1995b. Sampling for site-specific application, in P. C. Robert, R. H. Rust, and W. E. Larson, Ed., *Site-Specific Management for Agricultural Systems, 2nd Intl. Conf.,* ASA, CSSA, SSSA, Madison, WI, 535–551.

Freese, F. 1967. Elementary statistical methods for foresters, *Agric. Handb.* 317, U.S. Department of Agriculture Forest Service, Burgess Publishing Company, Minneapolis, MN.

Friendly, M. 1991. *SAS System for Statistical Graphics,* 1st ed., SAS Series in Statistical Applications, SAS Institute, Cary, NC.

Gauch, H. G. 1992. *Statistical Analysis of Regional Yield Trials. AMMI Analysis of Factorial Designs,* Elsevier, Amsterdam.

Gazey, W. J. and M.J. Staley. 1986. Population estimation from mark-recapture experiments using a sequential Bayes algorithm, *Ecology* 67:941–951.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. *Bayesian Data Analysis,* Chapman & Hall, London.

Giesler, R. and U. Lundström. 1993. Soil solution chemistry: effects of bulking soil samples, *Soil Sci. Soc. Am. J.* 57:1283–1288.

Gonick, L. and W. Smith. 1993. *The Cartoon Guide to Statistics,* HarperCollins, New York.

Gotway, C. A., R. B. Ferguson, G. W. Hergert, and T. A. Peterson. 1996. Comparison of kriging and inverse-distance methods for mapping soil parameters, *Soil Sci. Soc. Am. J.* 60:1237–1247.

Gotway, C. A., R. B. Ferguson, and G. W. Hergert. 1997. The effects of mapping and scale on variable rate fertilizer recommendations for corn, in P. C. Robert and W. E. Larson, Eds., *Site-Specific Management for Agricultural Systems, Third Intl. Conf.,* ASA, CSSA, SSSA, Madison, WI.

Hammond, R. B. and L. P. Pedigo. 1976. Sequential sampling plans for the green cloverworm in Iowa soybeans, *J. Econ. Entomol.* 69:181–185.

Hergert, G. W., R. B. Ferguson, and C. A. Shapiro. 1995a. Fertilizer Suggestions for Corn, University Nebraska NebGuide G74–174 (Revised).

Hergert, G. W., R. B. Ferguson, C. A. Shapiro, E. J. Penas, and F. B. Anderson. 1995b. Classical statistical and geostatistical analysis of soil nitrate-N spatial variability, in P. C. Robert, R. H. Rust, and W. E. Larson, Eds., *Site-Specific Management for Agricultural Systems, Second Intl. Conf.,* ASA, CSSA, SSSA, Madison, WI, 175–186.

Hergert, G. W., W. L. Pan, D. R. Huggins, J. H. Grove, and T. R. Peck. 1997. The adequacy of current fertilizer recommendations for site specific management, in F. J. Pierce and E. J. Sadler, Eds., *The state of Site-Specific Management for Agriculture*, ASA, CSSA, SSSA, Madison, WI, 283–300.

Higley, L. G. and D. J. Boethel. 1994. *Handbook of Soybean Insect Pests*, Entomological Society of America, Lanham, MD.

James, D. W. and K. L. Wells. 1990. *Soil sample collection and handling: technique based on source and degree of field variability*, in R. L. Westerman, Ed., *Soil Testing and Plant Analysis*, 3rd ed., SSSA, Madison, WI, 25–44.

Kaiser, L. 1983. Unbiased estimation in line-intercept sampling, *Biometrics* 39:965–976.

Kitchen, N. R., J. L. Havlin, and D. G. Westfall. 1990. Soil sampling under no-till banded phosphorus, *Soil Sci. Soc. Am. J.* 54:1661–1665.

King, J. R. 1980. Frugal Sampling Schemes, Technical and Engineering Aids for Management, Tamworth, NH.

Kogan, M. and D. C. Herzog, Eds., 1980. *Sampling Methods in Soybean Entomology*, Springer-Verlag, New York.

Little, T. M. and F. J. Hills. 1978. *Agricultural Experimentation. Design and Analysis*, John Wiley and Sons, New York.

Ludwig, J. A. and J. F. Reynolds. 1988. *Statistical Ecology. A Primer on Methods and Computing*, Wiley Interscience, New York.

Matheron, G. 1971. The theory of regionalized variables and its application, Cah. Cent. Morphol. Math. Fontainebleau 5. Centre de Geostatistique.

McKinion, J. M. 1992. Getting started: basics of modeling strategies, in J. L. Goodenough and J. M. McKinion, Eds., *Basics of Insect Modeling*, ASAE Monograph No. 10., 1–8.

McDonald, L. L. and B. F. J. Manly. 1989. Calibration of biased sampling procedures, in L. McDonald, B. Manly, J. Lockwood, and J. Logan, Eds., *Estimation and Analysis of Insect Populations*, Springer-Verlag, Berlin.

Miller, I. and J. E. Freund. 1977. *Probability and Statistics for Engineers*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Milliken, G. A. and D. E. Johnson. 1984. *Analysis of Messy Data*, Vol. 1, *Designed Experiments*, Van Nostrand Reinhold, New York.

Milliken, G. A. and D. E. Johnson. 1989. *Analysis of Messy Data*, Vol. 2, *Nonreplicated Experiments*, Van Nostrand Reinhold, New York.

Morgan, M. F. 1932. Microchemical Soil Tests, Connecticut Agric. Exp. Stn. Bull. 333.

Nyrop, J. P., R. E. Foster, and D. Onstad. 1986. Value of sample information in pest control decision making, *J. Econ. Entomol.* 79:1421–1429.

Parkin, T. B., J. J. Meisinger, S. T. Chester, J. L. Starr, and J. A. Robinson. 1988. Evaluation of statistical estimation methods for log normally distributed variables, *Soil Sci. Soc. Am. J.* 52:323–329.

Peck, T. R. and S. W. Melsted. 1967. Field sampling for soil testing, in M. Stelly, Ed., *Soil Testing and Plant Analysis*, Part 1: Soil Testing, SSSA, Madison, WI, 25–35.

Peck, T. R. and S. W. Melsted. 1973. Field sampling for soil testing, in L. M. Walsh and J. D. Beaton, Eds., *Soil Testing and Plant Analysis*, SSSA, Madison, WI, 67–75.

Peck, T. R. and P. M. Soltanpour. 1990. Principles of soil testing, in R. L. Westerman, Ed., *Soil Testing and Plant Analysis*, 3rd ed., SSSA, Madison, WI, 3–9.

Pedigo, L. P. and G. D. Buntin, Eds., 1993. *Handbook of Sampling Methods for Arthropods in Agriculture*, CRC Press. Boca Raton, FL.

Peterson, R. G. and L. D. Calvin. 1982. Sampling, in A. Klute, Ed., *Methods of Soil Analysis*, Part 1: *Agronomy*, 2nd ed., 9:33–51.

Pieters, E. P. 1978. Bibliography of Sequential Plans for Insects, *Bull. Entomol. Soc. Am.* 24(3):372–374.

Plant, R. E. and L. T. Wilson. 1985. A Bayesian method for sequential sampling and forecasting in agricultural pest management, *Biometrics* 41:203–214.

Randall, G. W. 1982. Strip tillage systems-fertilizer management, in *Farm Agric. Resources Management Conf. on Conservation Tillage*, Iowa State University Est. Publ. CE-1755.

Reed, J. F. and J. A. Rigney. 1947. Soil sampling from fields of uniform and non-uniform appearance and soil types, *J. Am. Soc. Agron.* 39:26–40.

Reuss, J. O., P. N. Soltapour, and A. E. Ludwick. 1977. Sampling distributions of nitrates in irrigated fields. *Agron. J.* 69:588–592.

Ruesink, W. G. 1980. Introduction to sampling theory, in M. Kogan and D. C. Herzog, Eds., *Sampling Methods in Soybean Entomology*, Springer-Verlag, New York.

Sabbe, W. E. and D. B. Marx. 1987. Soil sampling: spatial and temporal variability, in *Soil Testing: Sampling, Correlation, Calibration and Interpretation*, SSSA Spec. Publ. 21. SSSA, Madison, WI, 1–14.

Sawyer, J. E. 1994. Concepts of variable rate technology with considerations for fertilizer application, *J. Prod. Agric.* 7:195–201.

Schindler, E. 1996. *The Computer Speech Book*, Academic Press, Boston.

Schmitt, S. A. 1969. *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*, Addison-Wesley, Reading, MA, 400 pp.

Simon, J. L. and P. C. Bruce. 1993. *The New Biostatistics of Resampling*, Duxbury Press,

Spiegel, M. R. 1962. *Statistics. Schaum's Outline Series*, McGraw-Hill, New York.

Strayer, J., M. Shepard, and S. G. Turnipseed. 1977. Sequential sampling for management decisions on the velvetbean caterpillar on soybeans, *J. Ga. Entomol. Soc.* 12:220–227.

Thompson, S. K. 1992. *Sampling*, Wiley-Interscience, John Wiley and Sons, New York, 343 pp.

Truog, E. 1930. The determination of the readily available phosphorus in soils, *J. Am. Soc. Agron.* 22:874–882.

Waddill, V. H., B. M. Shepard, S. G. Turnipseed, and C. R. Carner. 1974. Sequential sampling plans for *Nabis* spp. and *Geocoris* spp. on soybeans, *Environ. Entomol.* 3:415–419.

Willers, J. L., D. L. Boykin, J. M. Hardin, T. L. Wagner, R. L. Olsen, and M. R. Williams. 1990a. A simulation study on the relationship between the abundance and spatial distribution of insects and selected sampling schemes, in *Proceedings, Applied Statistics in Agriculture*, Kansas State University, Manhattan, KS, 35–45.

Willers, J. L., R. L. Olson, M. R. Williams, and T. L. Wagner. 1990b. Developing a Bayesian approach for estimating the proportion of cotton plants at risk to insect attack, in *Proceedings, Beltwide Cotton Production Research Conferences*, Las Vegas, NV, 246.

Willers, J. L., S. R. Yatham, M. R. Williams, and D. C. Akins. 1992. Utilization of the line-intercept method to estimate the coverage, density, and average length of row skips in cotton and other row crops, in *Proceedings, Applied Statistics in Agriculture*, Kansas State University, Manhattan, KS, 48–59.

Williams, M. R., T. L. Wagner, and J. L. Willers. 1995. Revised Protocol for Scouting Arthropod Pests of Cotton in the Midsouth, Tech. Bull. 206, Mississippi Agricultural and Forestry Experiment Station, Mississippi State.

Wollenhaupt, N. C., D. J. Mulla, and C. A. Gotway Crawford. 1997. Soil sampling and interpolation techniques for mapping spatial variability of soil properties, in F. J. Pierce and E. J. Sadler, Eds., *The State of Site-Specific Management for Agriculture*, ASA, CSSA, SSSA, Madison, WI, 19–54.

Extension Service and Other Agricultural Publications

Lime Needes — Illustrated. Lime Increases Yield & Profit. Mississippi Cooperative Extension Service, Publ. 720.

Monitoring Soybeans for Insect Pests. Mississippi Cooperative Extension Service, Publ. 1498.

Soil Testing for the Farmer, Mississippi Cooperative Extension Service, Inf. Sheet 346.

Soybean Cyst Nematode, Mississippi Cooperative Extension Service, Publ. 1293.

Soybean Insect Control, Mississippi Cooperative Extension Service, Publ. 883.

Soybean Looper: Biology and Approaches for Improved Management, Mississippi Cooperative Extension Service, Inf. Sheet 1400.

Soybean Seedling Diseases, Mississippi Cooperative Extension Service, Inf. Sheet 1167.

Soybeans: Doublecropping Soybeans after Wheat in Mississippi, Mississippi Cooperative Extension Service, Publ. 1380.

Soybeans: Efficient Production Practices, Mississippi Cooperative Extension Service, Publ. 1559.

Soybeans: Plant Populations and Seeding Rates, Mississippi Cooperative Extension Service, Publ. 1194.

Soybeans: Planting Guidelines for Mississippi, Mississippi Cooperative Extension Service, Publ. 1289.

Stem Canker of Soybean, Mississippi Cooperative Extension Service, Publ. 1827.

Sterling, W. L., Ed., 1979. Economic Thresholds and Sampling of *Heliothis* Species on Cotton, Corn, Soybeans and Other Host Plants, Southern Cooperative Series Bull. 231. Department of Agricultural Communications, Texas A & M University, College Station, TX 77843.

Thom, W. O. and W. Sabbe, Eds., 1994. Soil Sampling Procedures for the Southern Region of the United States. Southern Cooperative Series Bull. 377, Kentucky Agricultural Experiment Station, Department of Agricultural Communications, University of Kentucky, Lexington, KY 40546.

Weeds of the Southern United States, Mississippi Cooperative Extension Service, No. 2500-8-75.