

Review & Comments

PROPOSED TECHNICAL NOTE

Statistical Methods

Statistical Methods

CONTENTS

- I. Introduction
- A. What Is Statistics?
 - B. Purpose Of This Technical Note
 - C. SAS: A Tool For Statistics
- II. Uses for Statistics in the SCS
- A. Data Collection (Sampling)
 - B. Data Analysis
 - C. Decision Making and Predictions
- III. Basic Statistics
- arithmetic*
- A. Mean
 - B. Standard Deviation and Variance
 - C. Coefficient Of Variation
 - D. Standard Error Of The Mean
 - E. Correlation Coefficient
 - F. Using SAS
- IV. Sampling Techniques
- A. Simple Random Sampling
 - B. Sample Size
- V. Confidence Limits (Reliability of The Sample)
- VI. Hypothesis Testing
- A. Testing The Mean
 - B. Comparing Two Groups Of Data
(Test Of Difference)
- VII. Linear Regression Analysis And Prediction
- A. Simple Regression
 - B. Multiple Regression

VIII. Appendix of Formulas

IX. Glossary

X. References

I. INTRODUCTION

What Is Statistics?

Webster defines statistics as: "the mathematics of the collection, organization, and interpretation of numerical data; especially the analysis of population characteristics by inference from sampling." (It is important to note that statistics includes organization and interpretation of data.) This definition has been "watered down" in the past because people have confused the science of statistics with the plural form of statistic, the latter being defined as the numerical data itself. Even though both are spelled the same, their meanings are quite different. To minimize confusion, a statistic or a group of statistics will be referred to as data in this paper. The science of statistics deals with:

1. Designing surveys
2. Collecting and summarizing data
3. Measuring the variation in survey data
4. Measuring the accuracy of estimates
5. Testing hypotheses about the population
6. Studying relationships among two or more variables

Statistics, then, boils down to the techniques used to obtain analytical measures, the methods for estimating their reliability, and the drawing of inferences from them.

Purpose of This Technical Note

The purpose of this technical note is to facilitate in the understanding of a variety of statistical techniques, their limitations, the assumptions behind them, and the interpretations that can be made from them. It is written under the assumption that SCS personnel will at some time be involved in a problem solving situation in which a working knowledge of basic statistical methods would be useful. The purpose is not to promote statistics as a panacea for all data problems, but to develop a guide for those who have an opportunity to apply statistical analysis in their work.

SAS: A Tool For Statistics

The Statistical Analysis System (SAS) is a computer software system originally developed for statistical needs. SAS is not a computer "language" in conventional terms as is FORTRAN or COBOL. SAS is a collection of prewritten programs which compute statistical measures and can be accessed by a few relatively simple commands.

One problem in writing a traditional statistical users guide is the ever-present mass of formulas needed throughout to develop analytical measures. Their presence often leaves the inquirer frustrated, discouraged, and ultimately opposed to any further use or study of statistics. But, by incorporating SAS into this paper, the pages of formulas can be replaced by a few SAS programs. Ultimately, the statistical calculations can be done easily on a computer without the need for memorizing formulas. The formulas are, however, included in an Appendix so that the user is able to study the complete derivations.

II. USES FOR STATISTICS IN THE SCS

No matter what activities are underway in the SCS, it's very likely they involve problem-solving in some form. The SCS technical specialist is often called into this process to interject his/her expertise whether it be in project work, conservation operations, or other SCS endeavors. The specialist uses the tools that he/she has learned in either formal or informal training as a means to answer questions and solve problems. One tool that can be useful in certain situations is statistics. There are basically three areas that the user of statistical methods is equipped to deal with:

1. Data collection (sampling)
2. Data analysis
3. Decision-making and predictions

Data Collection (Sampling)

One of the most fundamental steps in answering questions and solving problems is the collection of relevant data about the problem. People are constantly in pursuit of the "correct" information upon which to make day to day decisions. For most, the accuracy of information is only as important as the decision to be made. For example, in collecting facts to determine which brand of laundry detergent to purchase, one might ask a neighbor or a friend. Most, however, usually are content to learn by trial and error. And why not? The most risked if the choice is incorrect is a few dollars. But, when one sets out to purchase a new automobile, usually a number of owners are questioned, many dealers are consulted, and consumer periodicals are scoured to obtain data for the decision of whether or not to buy. After all, a new car is not a minor investment.

The use of information gathering in SCS is not unlike that of daily life. SCS personnel are constantly weighing whether or not additional information (obtained at some cost) will yield results which justify the effort and expense. Some problems encountered are too minor to justify a large expenditure for detailed data collection. On the other hand, some decisions involve thousands of dollars and years of development. It's only common sense to appropriate more effort to those types of decisions. But, where is the line drawn? How far should one go to insure the accuracy of data collection efforts? In most situations, it is impossible to collect information from every data source. For example, the SCS may need to find out the extent conservation tillage is being used in the United States. It would be impossible to contact and interview every farm operator. So a sample is taken from the total of all farmers. But, is the sample large enough to reflect the "true" information about all farmers? Conversely, funds been wasted by taking a larger sample than necessary? Statistical

methods can be used to help answer these questions. With the aid of statistics during data collection, sample size which will meet specific precision criteria for each individual situation can be determined. Chapter IV deals specifically with the use of statistics in data collection.

Data Analysis

When data is received from a sample, whether from a primary or a secondary source, confidence in that data is always a question. Using the conservation tillage example, it is important to know that the farmers surveyed are representative of all farmers. It is possible, through the use of statistical methods, to quantify "confidence" in the reliability of the sample data characterizing the total data. In fact, many of the basic statistical measures in Chapter III can be tested for the degree of reliability they possess. That is a unique characteristic of modern statistical methods.

Why is the ability to quantify confidence in statistical measurements so important? Imagine assigning two people to independently carry out the study of a particular problem. To make an intelligent decision, both people separately survey the problem, analyze the data, and report their results. Unfortunately, their results are very different. Person A subjectively decided to take a 10% sample. He averaged the data and reported the average (mean). Person B used statistics to determine the most economical sample size. When his sample was collected he, again using statistics, was able to report how well his sample represented the total. His average (mean) was reported as well but he also included the limits of confidence that should be assumed for his sample mean representing the actual mean.

The only thing that Person B did differently was to use statistical methods to determine sample size and to quantify confidence in his results. Now, which persons' analysis would be accepted? In most cases, a sample estimate which includes an indication of reliability is intuitively more acceptable than one which does not.

Decision Making and Predictions

The main purpose of most SCS studies is eventually to make some decision or to predict a future trend or event. It is possible to use statistics in both cases. Hypothesis testing is a method that can be used to compare an estimate or intelligent guess, to actual sampled data. For example, if information was obtained that the average soybean yield in Iowa was 33 bushels per acre, and the figure seemed low, a quick sample could developed from Ag census data or some other obtainable source. Using this sample and the techniques of hypothesis testing, a statistical test could be performed to see if the average yield was indeed 33 bushels per acre. (Numerical examples of hypothesis testing are given in Chapter VI.

Predicting the future is a job most SCS technical specialists are expected to perform to some extent. Statistics provides a method called regression which enables the user to predict future trends based on past data. For example, it is possible to estimate future yields based on past yields and associated levels of fertilizer, herbicide, etc. This type of relationship between inputs in agriculture and associated yields is commonly known as a production function. Why is it called a function? Because the end result of regression analysis is a mathematical equation (function) which can be used for prediction. This is one of the most powerful tools in statistics and probably the most widely used. Regression analysis is covered in Chapter VII.

To summarize, statistics combined with SAS gives the user:

1. A number of ways to sample and analyze data, (Chapters III, IV, VI, and VII).
2. Simple, time-saving means to quantify the reliability of the sample (Chapter V).
3. Techniques to compare estimates and data (Chapters III and VI).
4. A way to predict future trends based on past events (Chapter VII).

III. BASIC STATISTICS

The following statistics are used to calculate representative values that summarize much of the information available in a set of data. Not all of the statistics are of equal importance in every instance but they are presented so that the reader will be aware of their existence, uses, advantages, and disadvantages.

Arithmetic Mean

Probably the most well known and often used of these representative values is the arithmetic mean, usually referred to as the sample mean when the data is a sample. The mean is merely the arithmetic average of all the data values (See Appendix of Formulas) and is intended to represent the "typical" value. Its advantages include ease of computation, common usage, and use in algebraic manipulation. But, the major disadvantage is that it is unduly affected by extreme values and may, in fact, be far from representative of the data. It then is important to be able to diagnose the variability of the data as well as the mean.

Standard Deviation and Variance

Perhaps the most widely used measure of data variability is the standard deviation. The standard deviation characterizes dispersion about the mean and gives an indication whether most of the data is close to the mean or spread out. Variation about the mean is often spoken of in terms of variance rather than standard deviation. The variance is simply the square of the standard deviation.

An advantage of these measures of dispersion is that they can be used to formulate confidence limits and hypothesis tests which will be discussed later. The major disadvantage develops when comparing 2 sets of data. If one data set has a mean equal to 15 while the second data set mean is 1000, it becomes obvious (after studying the formula for computing the standard deviation in the Appendix of Formulas) that the standard deviation measurement of the data set with large values will be higher than that of the one with small values. Yet, the true variability of the data in the data set with large values may actually be less. Consider the following example:

<u>Data Set A</u>	<u>Data Set B</u>
2	974
20	1010
4	974
10	990
25	1020
14	1002
18	1006
4	978
30	1030
23	1016

The mean of data Set A is 15 and the mean of data Set B is 1000. The standard deviation is 9.78 for data set A and 20.09 for data set B. It would be inaccurate to conclude that data Set B is more variable than data Set A because the magnitude of the means is so different. In essence, a new measurement is needed for comparing variability of data sets with widely differing means.

Coefficient of Variation

The coefficient of variation (CV) is a measure of relative variation and is calculated by dividing the standard deviation by the mean, then multiplying by 100. This measurement expresses variability relative to the absolute magnitude of the mean. Using the previous example, the CV of data Sets A and B are 65.17 and 2.01 respectively. Thus, with the magnitude of the mean accounted for, data set A shows more variability than does data Set B, even though the standard deviation of A is less than that of B.

A second advantage of using the CV is that one is able to compare the variability of data sets that are in different units. The CV is independent of the unit of measurement. It is proper to compare the CV for wheat yields to the CV for soil loss or population.

Overall, if one is interested in analyzing the variability of more than one set of data through comparisons, it would be more meaningful to use the coefficient of variation rather than the standard deviation.

Standard Error of the Mean

In the previous example, Data Set A is a random sample which was taken from a population. If the random sample were to be repeated, the mean would probably not be 15. In fact, numerous samples from the same population would yield different means. Sampling will be explained in more detail in the next chapter but there is an important statistic that follows from the sampling procedure. The standard error of the mean is a calculation which indicates the variability in sample means much like the standard deviation indicates variability in individual sample observations.

It would be possible to estimate the variability of sample means by actually taking repeated samples and using the standard deviation formula. But, the standard error of the mean calculation makes it possible to gain the same information using the one initial sample. The main use of the standard error of the mean is in the development of confidence limits which will be discussed in Chapter V.

Correlation Coefficient

In many instances, a sample with more than one variable (characteristic) may be taken. For example, if a sample is taken of farmers who use conservation tillage, data on the farmers' ages and the farmers' education level might be collected. The correlation coefficient can be used to numerically describe any correlation (relationship) that exists between the two characteristics. The correlation coefficient ranges numerically between 1 and -1. A positive fraction close to one would suggest that older age is associated with higher education. A negative fraction close to -1 would indicate that older age is associated with less education. And finally, a fraction close to zero would indicate that the data does not support any strong correlation between a farmer's age and his education.

Correlation should not be thought of as cause and effect but merely as directional association. For example, economists have found a positive correlation between education and income. In general, people with more education earn higher incomes than do people with less education. But, this correlation does not give us any statistical evidence to show that higher education causes higher incomes. In fact, causation may even run the other direction as people with higher incomes may buy more education just as they buy more automobiles and more vacations. Thus, higher incomes may cause higher education levels! Do not fall into the trap of extending correlation to causation. Causation will be discussed in Chapter VII when the use of regression analysis is explored.

Using SAS

The Statistical Analysis System (SAS) can be used in calculating many statistical measures. A SAS program consists of basically 3 parts, Figure 1: Job Control Language ^{1/} (JCL) (records 1-7 and last), input data (records 8-35), and procedural statements (36 through last-1). The first 5 records in Figure 1 are accounting type statements and would vary depending on the user and location. The 6th and 7th records are included in the group of JCL records and initiate the SAS system. The last record (//) simply ends the program.

The example in Figure 1 contains hypothetical data from 25 farms, 600 acres in size, from two counties. It contains data on acres conservation tilled (ACRESMIN), age of operator (AGE), years of formal education (EDUC), and county (COUNTY).^{2/} Record 8 merely names the data set to be created, called CONSTILL in our example, while record 9 names the 4 variables from which data was collected. (These names will be referred to in the procedural records). Records 11 to 35 contain the data itself and the data in each record is arranged in the same order as the variable names in record 9. There are many ways to input data in SAS other than what is presented here. (See Chapter III of the SAS User's Guide: Basics, SAS Institute Inc., 1982.)

^{1/} Only the 8 JCL records begin in Column 1. The rest begin after Column 1.

^{2/} The dollar sign is used for data which is not numeric; in this case the county name.

Figure 1. Initial SAS Program using hypothetical conservation tillage data.

Record
number

```
1 //SCS17DAC JOB (XXXXXXXXXX,RJ114),'FWTX/SCS-SNTC-DOUG';
2 //CLASS=C, TIME=(00,20), MSGLEVEL = (1,1)
3 /*ROUTE PRINT RMT114
4 /*LOGONID XXXXX
5 /*PASSWORD XXXX
6 //EXEC SAS
7 //SAS.SYSIN DD *
8 DATA CONSTILL;
9 INPUT ACRESMIN AGE EDUC COUNTY $;
10 CARDS;
11 190 50 14 Cook
12 135 59 14 Haynes
13 275 39 16 Haynes
14 185 49 14 Haynes
15 340 39 12 Cook
16 575 32 16 Haynes
17 210 51 14 Cook
18 95 55 8 Haynes
19 55 67 8 Cook
20 210 28 12 Haynes
21 280 35 14 Haynes
22 0 68 8 Cook
23 120 55 12 Haynes
24 80 59 12 Cook
25 600 30 18 Cook
26 415 29 18 Haynes
27 0 62 8 Haynes
28 395 27 16 Haynes
29 480 31 16 Cook
30 180 52 12 Cook
31 60 63 6 Haynes
32 0 61 12 Haynes
33 108 61 12 Haynes
34 225 46 12 Cook
35 295 37 16 Cook
36 First Procedural Record
.
.
.
Last - 1 Last Procedural Record
Last //
```

The procedural records are those which are used to initiate the prewritten programs of SAS. These programs include a number of applications for statistics as will become evident through use of the conservation tillage example. Because the procedural records vary depending upon the specific program desired, Figure 1 displays only space for the records (records 36. . .Last-1). Specific procedural records will be supplied in conjunction with the specific program as each chapter warrants.

The most basic procedure that can be initiated is a procedure to print out the data, Figure 2. Using the program in Figure 1 with "PROC PRINT;" inserted in line 36, (starting in column 2), a list of the data will be printed out which is numbered and named. In the example, Figure 2, data on acres conservation tilled (ACRES MIN) age (AGE), formal education (EDUC), and county (COUNTY) are listed. Since the data is from 25 farms, 25 observations are numbered in the output.

Figure 2. Output from the Print procedure. Record used: PROC PRINT;

OBS	ACRESMIN	AGE	EDUC	COUNTY
1	190	50	14	Cook
2	135	59	14	Haynes
3	275	39	16	Haynes
4	185	49	14	Haynes
5	340	39	12	Cook
6	575	32	16	Haynes
7	210	51	14	Cook
8	95	55	8	Haynes
9	55	67	8	Cook
10	210	28	12	Haynes
11	280	35	14	Haynes
12	0	68	8	Cook
13	120	55	12	Haynes
14	80	59	12	Cook
15	600	30	18	Cook
16	415	29	18	Haynes
17	0	62	8	Haynes
18	395	27	16	Haynes
19	480	31	16	Cook
20	180	52	12	Cook
21	60	63	6	Haynes
22	0	61	12	Haynes
23	108	61	12	Haynes
24	225	46	12	Cook
25	295	37	16	Cook

Observation one represents a 50 year old farmer from Cook County with two years post high school education who conservation tills 190 of his 600 acres. Observation two represents a 59 year old farmer from Haynes County with two years post high school education who conservation tills 135 of his 600 acres, and so on.

There are two major uses for this procedure. The first is to supply the user with a hard copy of his/her data. The second is to allow the user to easily check his/her data for keypunching errors.

Another way to view the data is by plotting it on a graph. SAS has a procedure called Plot which can provide a two dimensional graph with very little user effort. Figures 3 and 4 show that two records inserted into the program of Figure 1 ^{1/} produce a two variable graph which is scaled and labeled. In Figure 3, acres conservation tilled (ACRESMIN) is plotted against the farmers' age (AGE). ACRESMIN is designated to the vertical axis because it is listed first in the record "PLOT ACRESMIN * AGE;". If AGE was first, it would be on the vertical axis. Acres conservation tilled and age are considered inversely related because the two sets of data move in opposite directions. That is, lower levels of conservation tillage are associated with higher ages while higher levels of conservation tillage are associated with lower ages. This inverse relationship between the two sets of data can be recognized in a graph by a general downward slope of the points.

Figure 4 is a graph of ACRESMIN and EDUC. These two sets of data indicate graphically that conservation tilled acres and farmers' education are directly related because higher levels of conservation tillage are associated with higher levels of education among the farmers sampled. This direct relationship is characterized by an upward slope of the data points.

Most of this chapter has been devoted to defining a few basic statistics which are useful in problem solving professions. Difficulties arise in the use of many of these measures because the calculations are time-consuming and sometimes complex, especially if the data sets being studied are very large (see Appendix of Formulas). But, from the development of SAS came a procedure called Means which calculates these measures for the user. With one procedural record, Figure 5, a number of useful measures are produced for each variable (characteristic) of the data set.

Using the conservation tillage example, Figure 5, the average farmer (out of 25 surveyed) was 47 years old with almost a year of post high school education who conservation tilled 220 of his 600 acres. The minimum and maximum values of the 3 variables give a feel for the range of values in the data. Acres which were conservation tilled ranged from 0 to 600; age ranged from 27 to 68; and education ranged from 6th grade to Masters Degree.

^{1/} These records should be inserted in spaces designated "Procedural records" in Figure 1.

PLOT OF ACRESMIN*AGE LEGEND: A = 1 OBS, B = 2 OBS, ETC.

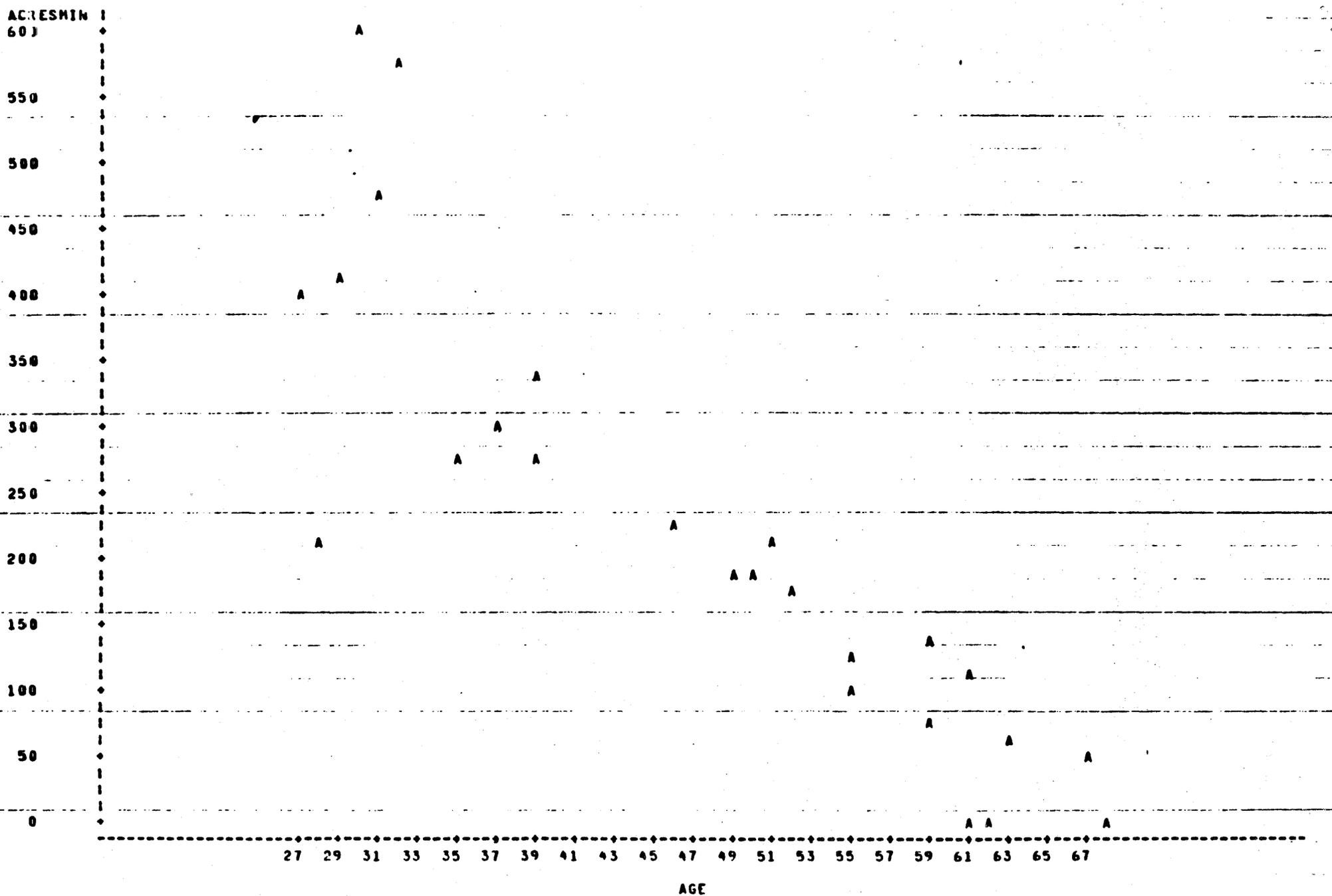


Figure 3. Output from the Plot procedure. Records used: PROC PLOT; PLOT ACRESMIN *AGE;

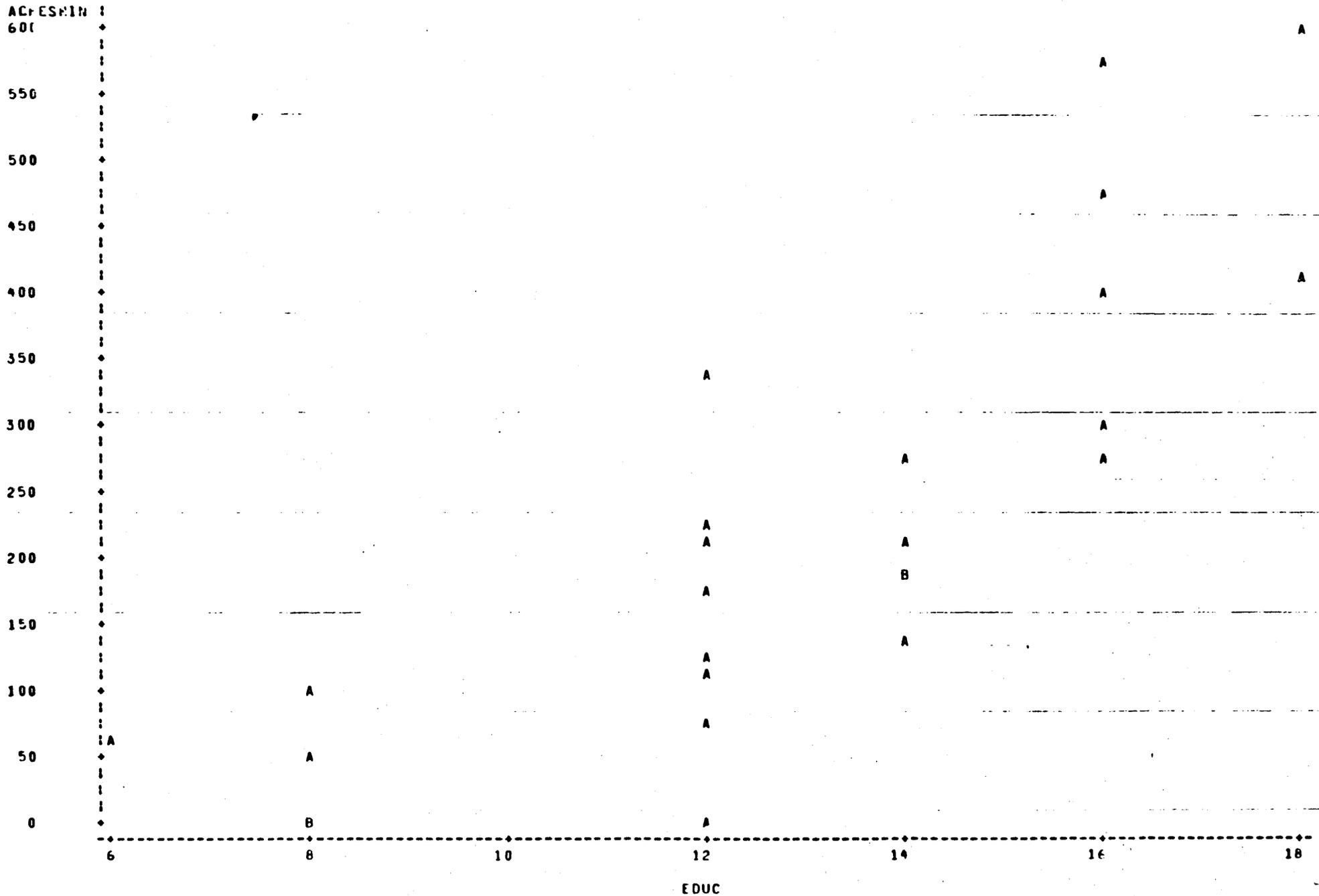


Figure Output from the Plot Procedure. Records used: PROC PLOT; PLOT ACRESMIN *EDUC;

The standard deviation and the variance (square of the standard deviation) indicate to what degree the data in general varies from the mean. The higher these measures are, the more dispersion exists in the data. These two measures are unit specific. That is if the mean of a variable is in acres, the standard deviation is in acres. The relative magnitude of the standard deviation is a function of the size of the mean. Therefore, it is meaningless to compare variability between two different variables, say ACRESMIN and AGE, as the magnitude of the means is so different (because they are in different units). It may be useful in some cases, however, to compare the standard deviation of conservation tilled acres in one sample to conservation tilled acres in some other sample. This is acceptable and would aid in choosing between two sets of data to study. As discovered previously in this chapter, a measure of relative variation (like the coefficient of variation) is used to compare variation between different variables. The coefficient of variation (CV) is calculated by the Means procedure of SAS and outputted for each variable, Figure 5.

Figure 5. Output from the Means procedure.
Record used: PROC MEANS;

<u>VARIABLE</u>	<u>N</u>	<u>MEAN</u>	<u>STANDARD DEVIATION</u>	<u>MINIMUM VALUE</u>	<u>MAXIMUM VALUE</u>
ACRESMIN	25	220.3200000	170.81806111	0.00000000	600.00000000
AGE	25	47.40000000	13.55236757	27.00000000	68.00000000
EDUC	25	12.80000000	3.26598632	6.00000000	18.00000000

<u>VARIABLE</u>	<u>STD ERROR OF MEAN</u>	<u>SUM</u>	<u>VARIANCE</u>	<u>C.V.</u>
ACRESMIN	34.16361222	5508.0000000	29178.810000	77.532
AGE	2.71047351	1185.0000000	183.666667	28.591
EDUC	0.65319726	320.0000000	10.666667	25.516

The standard error, which can be thought of as the standard deviation of sample means, is also supplied in the Means procedure. This measure is used extensively in the development of confidence intervals which will be discussed in Chapter V.

Correlation coefficients are developed in the correlation procedure of SAS as presented in Figure 6. (Some basic statistical measures presented in the Means procedure are also calculated here). In this example, the correlation coefficient between acres conservation tilled and age is $-.88$. The negative sign indicates an inverse relationship between the two variables, (as seen graphically in Figure 3), and the fraction close to one indicates a fairly strong relationship. Acres conservation tilled and education are positively correlated (directly related) as indicated by the coefficient of $.81$. As shown in Figure 4, this coefficient implies that higher education levels are associated with higher levels of conservation tillage. And finally, age and education are negatively correlated (inversely related) as older farmers are associated with less education according to the hypothetical data.

Figure 6. Output from the Correlation procedure.
Records used: PROC CORR;

<u>VARIABLE</u>	<u>N</u>	<u>MEAN</u>	<u>STD DEV</u>	<u>SUM</u>	<u>MINIMUM</u>	<u>MAXIMUM</u>
ACRESMIN	25	220.32	170.82	5508.00	0	600.00
AGE	25	47.40	13.55	1185.00	27.00	68.00
EDUC	25	12.80	3.27	320.00	6.00	18.00

CORRELATION COEFFICIENTS (N = 25)

	<u>ACRESMIN</u>	<u>AGE</u>	<u>EDUC</u>
ACRESMIN	1.00000	-0.88142	0.80688
AGE	-0.88142	1.00000	-0.77945
EDUC	0.80688	-0.77945	1.00000

IV. SAMPLING TECHNIQUES

Almost everyone in today's society is affected by sampling in one way or another. Polls are taken for public opinion, manufactured products are sampled for quality control, and consumers are surveyed to disclose their wants in market analyses. A carload of coal or grain is accepted or rejected based on analysis of a few pounds. Physicians make decisions about health based on a few drops of blood. Politicians are projected to win or lose based on results from a few precincts. The use of sampling is very widespread, yet the importance of sampling sometimes goes unnoticed. In SCS, day to day duties, are often fulfilled with studies or raw data that originated as a survey or sample of some kind.

A sample is basically a small collection of information from some larger aggregate, the population. The sample is collected and analyzed to make inferences about the total population. What makes this process more difficult is the presence of variation. If every farmer on earth was alike, a sample of 1 farmer would represent all farmers. Since this is not the case and members of a population are usually different, successive samples are usually different. Thus, the major task is to reach appropriate conclusions about the population in spite of sampling variation.

There are a number of different samples that are used. They are distinguished by the manner in which the sample is obtained, the number of variables recorded, and the purpose for drawing the sample. When considering the method of collection, two broad classes of sampling are possible; collection by judgement and collection by chance. Collection by chance, called random sampling, is preferred over judgemental collection because with random sampling information from the observed sample can be mathematically deduced based on the laws of probability. The purpose of randomness is to make certain these laws apply. There are no such laws which can apply to personal judgement.

SIMPLE RANDOM SAMPLING

If a sample is chosen so that each value in the population has an equal and independent chance of being collected, it is a simple random sample. A sample of this type can be obtained in a number of ways. Drawing marbles from an urn, tossing coins, and throwing dice are classical methods of obtaining randomness. The more modern method is to use a table of random numbers, Table 1. This table contains 10,000 digits jumbled in random fashion with 5 x 5 blocks to facilitate reading. There are 100 rows and 100 columns numbered from 00 to 99.

To illustrate the use of a table of random numbers, assume a list of 700 farmers in the two county area (Cook and Haynes), numbered from 1-700. Suppose a sample of 25 farmers from the population of 700 is to be drawn.^{1/}

^{1/} The determination of sample size will be examined later.

Refer to a table of random numbers such as Table 1 to draw a random sample and proceed through the following steps:

1. Flip a coin (or any other method) to choose which page of the table to begin with.
2. Without direction, bring a pencil point down on the page so as to hit a digit.
3. Read this digit and the next three to the right, e.g. 3478.
4. Let the first two digits (34) signify a row and the last two digits (78) signify a column.
5. Go to the point in the table (row 34, column 78) and read that digit (9) and the next two on the right. This number is 960.
6. Starting at this point, run down the column and record the numbers between 001 and 700. The numbers observed are 960, 807, 561, 431, 412, 821, etc. Record 561, 431, and 412, etc. because they are between 001 and 700, which is within the range of numbers assigned to the population of farmers. When the bottom of the page is reached, number 964, move to the next 3 columns to the right and start back up. Continue this process until 25 numbers between 001 and 700 are collected. The farmers that correspond to these 25 numbers constitute a random sample.

The method presented here is not sacred. It makes no difference whether movement is up or down, left or right. As long as the digits are collected in this general fashion of randomness, any contrived technique is permissible.

Sample Size

Before the random number table is of any use, it must be decided how large a sample is needed. As discussed earlier, too small a sample may lead to inaccurate assumptions about the population. Observing one farmer's conservation tillage methods does not tell much about the other 699 farmers in the area. Yet, it is not necessary to survey all 700 farmers either. This involves undue expense and time.

One method for determining adequate sample size uses the following equation:

$$n = \frac{tV}{E^2}$$

Table 1. Random number table, reproduced by permission from Snedecor and Cochran's "Statistical Methods" (ed. 7), Iowa State University Press, 1980.

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186	59391	58030	52098	82718	87024	82848	04190	96574	90464	29065
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927	99567	76364	77204	04615	27062	96621	43918	01896	83991	51141
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345	10363	97518	51400	25670	98342	61891	27101	37855	06235	33316
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455	86859	19558	64432	16706	99612	59798	32803	67708	15297	28612
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577	11258	24591	36863	55368	31721	94335	34936	02566	80972	08188
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976	95068	88628	35911	14530	33020	80428	39936	31855	34334	64865
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305	54463	47237	73800	91017	36239	71824	83671	39892	60518	37092
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779	16874	62677	57412	13215	31389	62233	80827	73917	82802	84420
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279	92494	63157	76593	91316	03505	72389	96363	52887	01087	66091
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231	15669	56689	35682	40844	53256	81872	35213	09840	34471	74441
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083	99116	75486	84989	23476	52967	67104	39495	39100	17217	74073
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113	15696	10703	65178	90637	63110	17622	53988	71087	84148	11670
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182	97720	15369	51269	69620	03388	13699	33423	67453	43269	56720
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551	11666	13841	71681	98000	35979	39719	81899	07449	47985	46967
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491	71628	73130	78783	75691	41632	09847	61547	18707	85489	69944
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770	40501	51089	99943	91843	41995	88931	73631	69361	05375	15417
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104	22518	55576	98215	82068	10798	86211	36584	67466	69373	40054
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771	75112	30485	62173	02132	14878	92879	22281	16783	86352	00077
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106	80327	02671	98191	84342	90813	49268	95441	15496	20168	09271
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067	60251	45548	02146	05597	48228	81366	34598	72856	66762	17002
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436	57430	82270	10421	05540	43648	75888	66049	21511	47676	33444
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620	73528	39559	34434	88596	54086	71693	43132	14414	79949	85193
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337	25991	65959	70769	64721	86413	33475	42740	06175	82758	66248
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119	78388	16638	09134	59880	63806	48472	39318	35434	24057	74739
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037	12477	09965	96657	57994	59439	76330	24596	77515	09577	91871
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546	83266	32883	42451	15579	38155	29793	40914	65990	16255	17777
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848	76970	80876	10237	39515	79152	74798	39357	09054	73579	92359
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700	37074	65198	44785	68624	98336	84481	97610	78735	46703	98265
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283	83712	06514	30101	78295	54656	85417	43189	60048	72781	72606
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163	20287	56862	69727	94443	64936	08366	27227	01558	50326	59566
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880	74261	32592	86538	27041	65172	85532	07571	80609	39285	65340
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631	64081	49863	08478	96001	18888	14810	70545	89755	59064	07210
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655	05617	75818	47750	67814	29575	10526	66192	44464	27058	40467
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864	26793	74951	95466	74307	13330	42664	85515	20632	05497	33625
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859	65988	72850	48737	54719	52056	01526	03845	35067	03134	70322
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567	27366	42271	44300	73399	21105	03280	73457	43093	05192	48657
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557	56760	10909	98147	34736	33863	95256	12731	66598	50771	83665
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887	72880	43338	93643	58904	59543	23943	11231	83268	65938	81581
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258	77888	38100	03062	58103	47961	83841	25878	23746	55903	44115
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716	28440	07819	21580	51459	47971	29882	13990	29226	23608	15873
40	79152	53829	77250	20190	56535	18760	69942	77448	33278	48805	63525	94441	77033	12147	51054	49955	58312	76923	96071	05813
41	44560	38750	83635	56540	64900	42912	13953	79149	18710	68618	47606	93410	16359	89033	89696	47231	64498	31776	05383	39902
42	68328	83378	63369	71381	39564	05615	42451	64559	97501	65747	52669	45030	96279	14709	52372	87832	02735	50803	72744	88208
43	46939	38689	58625	08342	30459	85863	20781	09284	26333	91777	16738	60159	07425	62369	07515	82721	37875	71153	21315	00132
44	83544	86141	15707	96256	23068	13782	08467	89469	93842	55349	59348	11695	45751	15865	74739	05572	32688	20271	65128	14551
45	91621	00881	04900	54224	46177	55309	17852	27491	89415	23466	12900	71775	29845	60774	94924	21810	38636	33717	67598	82521
46	91896	67126	04151	03795	59077	11848	12630	98375	52068	60142	75086	23537	49939	33595	13484	97588	28617	17979	70749	35234
47	55751	62515	21108	80830	02263	29303	37204	96926	30506	09808	99495	51434	29181	09993	38190	42553	68922	52125	91077	40197
48	85156	87689	95493	88842	00664	55017	55539	17771	69448	87530	26075	31671	45386	36583	93459	48599	52022	41330	60651	91321
49	07521	56898	12236	60277	39102	62315	12239	07105	11844	01117	13636	93596	23377	51133	95126	61496	42474	45141	46660	42338

Table 1. (continued)

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99
50	64249	63664	39652	40646	97306	31741	07294	84149	46797	82487	32847	31282	03345	89593	69214	70381	78285	20054	91018	16742
51	26538	44249	04050	48174	65570	44072	40192	51153	11397	58212	16916	00041	30236	55023	14253	76582	12092	86533	92426	37655
52	05845	00512	78630	55328	18116	69296	91705	86224	29503	57071	66176	34047	21005	27137	03191	48970	64625	22394	39622	79085
53	74897	68373	67359	51014	33510	83048	17056	72506	82949	54600	46299	13335	12180	16861	38043	59292	62675	63631	37020	78195
54	20872	54570	35017	88132	25730	22626	86723	91691	13191	77212	22847	47839	45385	23289	47526	54098	45683	55849	51575	64689
55	31432	96156	89177	75541	81355	24480	77243	76690	42507	84362	41851	54160	92320	69936	34803	92479	33399	71160	64777	83378
56	66890	61505	01240	00660	05873	13568	76082	79172	57913	93448	28444	59497	91586	95917	68553	28639	06455	34174	11130	91994
57	48194	57790	79970	33106	86904	48119	52503	24130	72824	21627	47520	62378	98855	83174	13088	16561	68559	26679	06238	51254
58	11303	87118	81471	52936	08555	28420	49416	44448	04269	27029	34978	63271	13142	82681	05271	08822	06490	44984	49307	62717
59	54374	57325	16947	45356	78371	10563	97191	53798	12693	27928	37404	80416	69035	92980	49486	74378	75610	74976	70056	15478
60	64852	34421	61046	90849	13966	39810	42699	21753	76192	10508	32400	65482	52099	53676	74648	94148	65095	69597	52771	71551
61	16309	20384	09491	91588	97720	89846	30376	76970	23063	35894	89262	86332	51718	70663	11623	29834	79820	73002	84886	03591
62	42587	37065	24526	72602	57589	98131	37292	05967	26002	51945	68866	09127	98021	03871	27789	58444	44832	36505	40672	30180
63	40177	98590	97161	41682	84533	67588	62036	49967	01990	72308	90814	14833	08759	74645	05046	94056	99094	65091	32663	73040
64	82309	76128	93965	26743	24141	04838	40254	26065	07938	76236	19192	82756	20553	58446	55376	88914	75096	26119	83898	43816
65	79788	68243	59732	04257	27084	14743	17520	95401	55811	76099	77585	52593	56612	95766	10019	29531	73064	20953	53523	58136
66	40538	79000	89559	25026	42274	23489	34502	75508	06059	86682	23757	16364	05096	03192	62386	45389	85332	18877	55710	96459
67	64016	73598	18609	73150	62463	33102	45205	87440	96767	67042	45989	96257	23850	26216	23309	21526	07425	50254	19455	29315
68	49767	12691	17903	93871	99721	79109	09425	26904	07419	76013	92970	94243	07316	41467	64837	52406	25225	51553	31220	14032
69	76974	55108	29795	08404	82684	00497	51126	79935	57450	55671	74346	59596	40088	98176	17896	86900	20249	77753	19099	48885
70	23854	08480	85983	96025	50117	64610	99425	62291	86943	21541	87646	41309	27636	45153	29988	94770	07255	70908	05340	99751
71	68973	70551	25098	78033	98573	79848	31778	29555	61446	23037	50099	71038	45146	06146	55211	99429	43169	66259	97786	59180
72	36444	93600	65350	14971	25325	00427	52073	64280	18847	24768	10127	46900	64984	75348	04115	33624	68774	60013	35515	62556
73	03003	87800	07391	11594	21196	00781	32550	57158	58887	73041	67995	81977	18984	64091	02785	27762	42529	97144	80407	64524
74	17540	26188	36647	78386	04558	61463	57842	90382	77019	24210	26304	80217	84934	82657	69291	35397	98714	35104	08187	48109
75	38916	55809	47982	41968	69760	79422	80154	91486	19180	15100	81994	41070	56642	64091	31229	02595	13513	45148	78722	30144
76	64288	19843	69122	42502	48508	28820	59933	72998	99942	10515	59537	34662	79631	89403	65212	09975	06118	86197	58208	16162
77	86809	51564	38040	39418	49915	19000	58050	16899	79952	57849	51228	10937	62396	81460	47331	91403	95007	06047	16846	64809
78	99800	99566	14742	05028	30033	94889	53381	23656	75787	59223	31089	37995	29577	07828	42272	54016	21950	86192	99046	84864
79	92345	31890	95712	08279	91794	94068	49337	88674	53555	12267	38207	97938	93459	75174	79460	55436	67206	87644	21296	43395
80	90363	65162	32245	82279	79256	80834	06088	99462	56705	06118	88666	31142	09474	89712	63153	62333	42212	06140	42594	43671
81	64437	32242	48431	04835	39070	59702	31508	60935	22390	52246	53365	56134	67582	92557	89520	33452	05134	70628	27612	33736
82	91714	53662	28373	34333	55791	74758	51144	18827	10704	76803	89807	74530	38004	90102	11693	90257	05500	79920	62700	43325
83	20902	17646	31391	31459	33315	03444	55743	74701	58851	27427	18682	81038	85662	90915	91631	22223	91588	80774	07716	12548
84	12217	86007	70371	52281	14510	76094	96579	54853	78339	20839	63571	32579	63942	25371	09234	94592	98475	76884	37635	33608
85	45177	02863	42307	53571	22532	74921	17735	42201	80540	54721	68927	56492	67799	95398	77642	54913	91853	08424	81450	76229
86	28325	90814	08804	52746	47913	54577	47525	77705	95330	21866	56401	63186	39389	88798	31356	89235	97036	32341	33292	73757
87	29019	28776	56116	54791	64604	08815	46049	71186	34650	14994	24333	95603	02359	72942	46287	95382	08452	62862	97869	71775
88	84979	81353	56219	67062	26146	82567	33122	14124	46240	92973	17025	84202	95199	62272	06366	16175	97577	99304	41587	03686
89	50371	26347	48513	63915	11158	25563	91915	18431	92978	11591	02804	08253	52133	20224	68034	50865	57868	22343	55111	03607
90	53422	06825	69711	67950	64716	18003	49581	45378	99878	61130	08298	03879	20995	19850	73090	13191	18963	82244	78479	99121
91	67453	35651	89316	41620	32048	70225	47597	33137	31443	51445	59883	01785	82403	96062	03785	03488	12970	64896	38336	30030
92	07294	85353	74819	23445	68237	07202	99515	62282	53809	26685	46982	06682	62864	91837	74021	89094	39952	64158	79614	78235
93	79544	00302	45338	16015	66613	88968	14595	63836	77716	79596	31121	47266	07661	02051	67599	24471	69843	83696	71402	76287
94	64144	85442	82060	46471	24162	39500	87351	36637	42833	71875	97867	56641	63416	17577	30161	87320	37752	73276	48969	41915
95	90919	11883	58318	00042	52402	28210	34075	33272	00840	73268	57364	86746	08415	14621	49430	22311	15836	72492	49372	44103
96	06670	57353	86275	92276	77591	46924	60839	55437	03183	13191	09559	26263	69511	28064	75999	44540	13337	10918	79846	54809
97	36634	93976	52062	83678	41256	60948	18685	48992	19462	96062	53873	55571	00608	42661	91332	63956	74087	59008	47493	99581
98	75101	72891	85745	67106	26010	62107	60885	37503	55461	71213	35531	19162	86406	05299	77511	24311	57257	22826	77555	05941
99	05112	71222	72654	51583	05228	62056	57390	42746	39272	96659	28229	88629	25695	94932	30721	16197	78742	34974	97528	45447

where:

n = size of the sample (what's being solving for)
t = confidence constant: the level of confidence required in
the sample mean; 4 for 95% confidence and 6.76 for 99% confidence
E = acceptable error from the true mean
V = variance

For example, if there is a need to estimate the average use of conservation tillage of farmers in Cook and Haynes Counties plus or minus 50 acres with 95% confidence, t would equal 4 and E would equal 50. The choice of values for t and E depends on the degree of precision wanted. The ability to estimate t and E comes with a feel for the population being sampled and experience in sampling.

The variance of the sample (V) is calculated from the sample data. So, how can an estimate of V be provided to calculate sample size if no data has been collected yet? The best way to obtain V is to take a pre-sample, calculate that variance, and use it in finding the sample size. But, pre-sampling is normally costly and time consuming. Thus, the following formula was developed to give a rough estimate of V without pre-sampling:

$$V = (R/4)^2$$

where R = range of data expected.

In the conservation tillage example, the population consists of 700 farms which are 600 acres in size. The main variable of interest is the acreage which is conservation tilled. It's quite obvious that this amount could range from 0-600 since a farmer could either conservation till all or none of his farm. Using the V formula with $R = 600 - 0 = 600$, $V = (600 \div 4)^2 = 22,500$.

With a value for V, sample size n can be calculated with a confidence constant of 4 (95% confidence) and an acceptable error of 50 as follows:

$$n = \frac{4 (22,500)}{50^2} = 36$$

To demand a 99% degree of confidence and require estimates to be within 40 acres of the true mean would necessitate a sample of 95 farmers rather than 36:

$$n = \frac{6.76 (22,500)}{40^2} = 95$$

Generally as estimated variance and required degree of confidence increases and acceptable error decreases, sample size must increase.

V. CONFIDENCE LIMITS

(Reliability of the Sample)

A point estimate from a sample, such as the sample mean, is usually not very meaningful by itself. It is almost certain to be less than exact (due to sampling variation) and it gives absolutely no indication of how erroneous it is. Any sample is subject to sampling error due to variability in the population. This variation can be realized by the comparison of multiple samples from the same population. For example, five different samples of age from the 700 farmers in Cook and Haynes counties would yield five different sample means. These different sample means reflect the variation which is inherent in the population of 700 farmers. Generally, only one sample is taken so the mean of the sample should be accompanied by some interval, together with some measure of assurance that the true population mean lies within the interval.

The statistical method for indicating reliability of a sample involves the use of confidence limits. Confidence limits for the mean of a sample express the true population mean as an interval estimate (between 2 values) with a certain probability. For example, the upper and lower confidence limits for the mean of a sample of age from the population of 700 farmers could be expressed as follows:

37 population mean 44 with 90% confidence.

This statement suggests that true population mean can be predicted to fall between 37 and 44 in 9 out of 10 samples. Conversely, the interval will not contain the population mean in 1 out of 10 samples. The most common percents of confidence used in this method are 90%, 95%, and 99%. Since 90% confidence is the same as 10% error and 95% confidence is the same as 5% error, etc., the most common probabilities of error are 10%, 5%, and 1% respectively. An interval of reliability can be calculated for the mean of a sample in the following manner:

Confidence Limits = mean \pm (t) (standard error of the mean)

SAS will compute the mean and the standard error of the mean in the MEANS procedure, Figure 5. The only thing left to determine is "t". "t" is a confidence coefficient which is based on "degrees of freedom" (in general, degrees of freedom equal the number of observations in the sample minus 1) and the percent confidence chosen. A matrix of "t" values is given in Table 2. Using the age of farmers example, sample size is 25 so degrees of freedom is 24. If 95% confidence is chosen as the level of reliability

TABLE 2. "t" table*

Degrees of Freedom	Probability of Error and (% Confidence)			
	.50 (50%)	.10 (90%)	.05 (95%)	.01 (99%)
1	1.00	6.34	12.71	63.66
2	.816	2.92	4.30	9.92
3	.765	2.35	3.18	5.84
4	.741	2.13	2.78	4.60
5	.727	2.02	2.57	4.03
6	.718	1.94	2.45	3.71
7	.711	1.90	2.36	3.50
8	.706	1.86	2.31	3.36
9	.703	1.83	2.26	3.25
10	.700	1.81	2.23	3.17
11	.697	1.80	2.20	3.11
12	.695	1.78	2.18	3.06
13	.694	1.77	2.16	3.01
14	.692	1.76	2.14	2.98
15	.691	1.75	2.13	2.95
20	.687	1.72	2.09	2.84
30	.683	1.70	2.04	2.75
40	.681	1.68	2.04	2.71
50	.679	1.68	2.02	2.68
75	.678	1.67	2.00	2.65
100	.677	1.66	1.98	2.63
125	.676	1.66	1.98	2.62
150	.676	1.66	1.98	2.61
200	.675	1.65	1.97	2.60
300	.675	1.65	1.97	2.59
400	.675	1.65	1.97	2.59
500	.674	1.65	1.96	2.59
1000	.674	1.65	1.96	2.58
∞	.674	1.64	1.96	2.58

*Parts of this table have been reproduced by permission from Snedecor and Cochran's "Statistical Methods" (ed. 7), Iowa State University Press, 1980.

required, "t" can be obtained from moving down the .05 probability of error (95% confidence) column. The "t" value for 20 degrees of freedom is 2.09 and the "t" for 30 degrees of freedom is 2.04. To find the "t" for 24 degrees of freedom, interpolate and find 2.07 as "t". From Figure 5 the mean for AGE is 47.4 and the standard error is 2.71. Substituting these values into the equation yields:

$$\text{Confidence Limits} = 47.4 \pm (2.07) (2.71) = 47.4 \pm 5.6$$

$$\text{Confidence Limits} = 41.8 \text{ and } 53.0$$

Thus, one can be 95% sure that the average age of the 700 farmers falls between 41.8 and 53.0 given the mean of the 25 farmer sample.

For another example, determine the 90% confidence limits for the mean of the education level of the farmers in the conservation tillage example. Figure 5 shows that SAS has computed the mean of the 25 farmer sample to be 12.8 years of education with a standard error of .65. Looking under the .10 (90%) column of the "t" table, Table 2, interpolate between 1.72 and 1.70 and find "t" equal to 1.71. Substituting these values into the equation:

$$\text{Confidence Limits} = 12.8 \pm (1.71) (.65) = 12.8 \pm 1.1$$

$$\text{Confidence Limits} = 11.7 \text{ and } 13.9$$

Thus, one can be 90% sure that the average education level of the 700 farmers falls between the near high school graduate level and the two year college level given the mean of the 25 farmer sample.

VI. HYPOTHESIS TESTING

A hypothesis can be defined as tentative theory or supposition. Everyone hypothesizes from time to time observations are made. For example, the following could be taken as hypotheses: (1) the average height of American adult males is 5 feet, 9 inches (2) the soybean yield in watershed x is 35 bushels per acre (3) the average row crop farmer in the Midwest conservation tills half his cropland. These three hypotheses are statistical hypotheses because they are statements about a statistical population; specifically they are statements about a certain variable (characteristic) in a statistical population.

It is often desirable to test if such hypotheses are valid. To do this, an appropriate sample is taken and the hypothesis is accepted or rejected based on the results of statistical tests. This chapter deals with two such statistical tests.

TESTING THE MEAN

The average 600 acre farmer in Cook and Haynes Counties is 55 years old, has an eight grade education, and conservation tills 200 of his 600 acres. This is a hypothesis. In fact, it is actually three separate hypotheses about the same 700 farmer population. How would one go about testing whether these estimates are accurate? One way would be to interview all 700 farmers. Although very thorough, this method involves extensive time and money. Too extensive for the resources of most government agencies. There is another option. It involves: (1) interviewing an appropriate sized random sample of the 700 farmers; (2) gathering information about age, education, and degree of conservation tillage; (3) calculating the sample mean for each of the 3 variables; and (4) using hypothesis testing methods to compare the sampled means to the original hypotheses.

Steps 1, 2, and 3 have already been completed (using hypothetical data). The results of these steps are recorded in Figure 5 where SAS has helped calculate the sample means. Step 4 involves the use and comparison of two "t" statistics; the "table t" found in Table 2 (just as was done for confidence intervals in Chapter V); and the "calculated t" which is calculated using information from the sample.

As was done to calculate confidence intervals, the percent confidence required of the test must be determined. Using this value, plus the degrees of freedom, the correct table t can be found in Table 2. In this case the sample size is 25 so degrees of freedom are one less than that or 24. If 90% confidence is required in the hypothesis test, the next step is to interpolate between 20 and 30 degrees of freedom in Table 2 under the .10 (90%) column and find that the table t is equal to ~~2.07~~.

1.71

The equation used to find the calculated t is:

calculated t = (sample mean - hypothesized mean) / standard error of the sample mean

In the MEANS procedure, SAS calculates the sample mean and the standard error of the sample mean as seen in Figure 5. To calculate the calculated t for age subtract the hypothesized mean (55) from the sample mean (47) which gives - 8. The absolute value (8) is then divided by the standard error of the mean (2.71) to yield a calculated t of 3.86:

$$\text{calculated t for age} = |47-55| \div 2.71=3.86$$

The comparison rule for table t and calculated t is as follows:

If calculated t exceeds table t reject the hypothesis; if not, accept the hypothesis.

The age hypothesis stated that the 700 farmers' ^(1.71) average age is 55. The calculated t (3.86) is greater than the table t ~~(2.07)~~ so this hypothesis is rejected. (The sample tends to show the average age is less than 55.) The test is done with the understanding that the decision to reject is done so with 90% confidence, thus a 10% margin of error is accepted.

To test the second hypothesis that the farmers' average education level is an eighth grade education, the same steps are followed using data from the education information in the sample, Figure 5. If the 90% level of confidence is satisfactory, the table t would be the same as in the age example, ~~2.07~~. The calculated t in this example would be:

$$\text{calculated t for education} = |12.8-8| \div .65 = 7.38$$

In the equation, 12.8 is the sample mean from SAS (Figure 5), 8 is the hypothesized mean, and .65 is the standard error of the mean from SAS (Figure 5). The calculated t (7.38) is larger than the table t ~~(2.07)~~ so the hypothesis that the average farmer in the population of 700 has an eighth grade education is rejected. (The sample would tend to show they have higher than an eighth grade education on the average.)

The third test is on the hypothesis that the average farmer conservation tills 200 of his 600 acre farm. Assuming a 90% confidence level again, the table t is ~~2.07~~. The calculated t is .59 using the output from SAS in Figure 5 under the conservation till variable, ACRES MIN:

$$\text{calculated t for conservation tilled acres} = |(220-200)| \div 34.16 = .59$$

The calculated t (.59) does not exceed the table t ~~(2.07)~~ so the hypothesis that the average farmer conservation tills 200 acres is accepted.

To summarize, statistical procedures were used to test 3 different hypotheses about 700 farmers in Cook and Haynes Counties. The procedure involved comparison of a value (calculated t) which comes from the combined results of the sample and the hypothesis to a value which incorporates sample size and an acceptable level of precision (table t).

The acceptance of a hypothesis reveals that the sample mean itself is no more accurate than the hypothesized value. (Although the methods used to obtain the sample mean may be more defensible.) The rejection of a hypothesis shows that the sample mean is statistically more accurate than the hypothesized mean although confidence in this decision is only as high as the level of confidence used to obtain the table t .

Comparing Two groups of Data (Test of Difference)

One of the most often used statistical hypothesis tests is the Test of Difference. The differences being investigated are those between 2 means, and the hypothesis being tested states that the two means are equal. Uses for this test could include comparing the rate of gain of hogs on 2 different rations, comparing math scores of males vs. females, or testing to see if the 300 farmers in Cook County vary from the 400 in Haynes County in terms of age, education, and use of conservation tillage. The actual hypothesis would state that there are no differences in rate of gain between rations; that there are no differences in average scores between males and females; and that there are no differences between average age, education, or use of conservation tillage between the 2 counties, respectively.

To test the 3 hypotheses that there is no difference between average use of conservation tillage, age, and education between farmers sampled in Cook and Haynes Counties, a "calculated t " and a "table t " are compared to determine whether each of the 3 hypotheses are accepted or rejected just as was done for the "Testing the Mean" procedure. However, the calculated t formula for the test of difference is much more complicated than the one for testing the mean (see Appendix of Formulas). Thus a SAS procedure TTEST has been developed to arrive at the calculated t for the Test of Difference.

To use the TTEST procedure the initial program, Figure 1, is run with 2 procedural records inserted in lines 36 and 37. "PROC TTEST;" initiates the prewritten SAS program which develops a calculated t for the variable in each hypothesis (acres conservation tilled, age and education). "CLASS COUNTY;" indicates the variable (COUNTY) upon which comparisons will be made.

Figure 7 illustrates the output from the TTEST procedure. The basic statistical measures are supplied for each variable and divided into the 2 counties. Calculated t and degrees of freedom are also supplied for use in the Test of Difference. The decision rule for the Test of Difference is: If the calculated t is larger than the table t , reject the hypothesis. (This is the identical rule used in Testing the Mean.)

For the hypothesis that the average use of conservation tillage is the same in Cook and Haynes Counties the calculated t is .538, Figure 7. Figure 7 also shows there are 23 degrees of freedom in this example. If 95% confidence is required in the test, the .05 (95%) column of Table 2 supplies the table t . Moving down the .05 (95%) column to 23 degrees of freedom an interpolation must be made between 2.04 and 2.09 yielding 2.06 as the table t . The calculated t is not larger than the table t so the hypothesis is calculated. Using the sample of 11 farmers from Cook County and 14 farmers from Haynes County plus the Test of Difference it can be concluded that the population of 300 farmers in Cook County conservation till, on average, the same proportions of their farms as do the 400 farmers in Haynes County.

The hypothesis which states that the average age of the 300 farmers in Cook County does not differ from the average age of the 400 farmers in Haynes County can be tested as well. The calculated t is .251, Figure 7, and if a 95% confidence was assumed again, the table t would again be 2.06. Since the calculated t is not larger than the table t the original hypothesis about the average age of the farmers is accepted. Using the same procedure it would also be concluded that the average education level between counties does not differ statistically for the total populations since the calculated t (.140) is less than the tabled t (2.06). ^{1/}

The Test of Difference is a useful technique in statistics. Many comparisons of large populations can be made using this test along with sampling techniques. This particular hypothesis testing procedure is enhanced since the calculations can be done easily using SAS.

^{1/} It is just by chance that all 3 example hypotheses were accepted, since the conservation tillage example used throughout this publication comes from hypothetical data. In real life it is probable that most counties are more diverse than those in this example.

Figure 7. Relevant output from the TTEST procedure.
Records used: PROC TTEST; CLASS COUNTY;

<u>County</u>	<u>N</u>	<u>Mean</u>	<u>Standard Deviation</u>	<u>Standard Error</u>	<u>Minimum</u>	<u>Maximum</u>	<u>t</u>	<u>Degrees of Freedom</u>
ACRESMIN								
COOK	11	241.36	180.24	54.34	0	600		
HAYNES	14	203.79	167.96	44.89	0	575	.538	23
AGE								
COOK	11	48.18	13.15	3.96	30	68		
HAYNES	14	46.79	14.32	3.83	27	63	.251	23
EDUC								
COOK	11	12.91	3.14	.95	8	18		
HAYNES	14	12.71	3.47	.93	6	18	.140	23

VII. Linear Regression Analysis and Prediction

In the previous chapters the problems considered involved only a single variable of a population at one time. Confidence limits were constructed around the mean of one variable. Hypothesis tests were developed for a single variable. The observation in a sample may have been compared in groups but generally this comparison was based on only one measurement or variable per comparison. In this chapter attention is turned to statistical influences based on two or more variables of each member of the sample. For example, more adequate judgements about a farmers' use of conservation tillage can be made if characteristics that may affect this use, like his age or education level can be studied simultaneously.

Linear regression analysis is concerned with the relationship of 2 or more variables. More specifically, it enables a user to determine to what degree one variable is affected by the others. In the conservation tillage example, farmers' use of conservation tillage may depend to some degree on their age and their education level. Linear regression analysis can be employed to mathematically and statistically describe the relationship of age and education to farmers' use of conservation tillage.

The major component of linear regression analysis is the linear regression model. This model may vary from application to application, but it can be expressed in general as: $Y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$

where: Y= variable to be explained called the dependent variable, eg. use of conservation tillage. (data from a sample)
B₀ = intercept (solved for using SAS)
X_i = variable or variables, used to explain Y, called the independent variables, eg. age, education. (data from a sample)
B_i = unknown parameters (solved for using SAS)
n_i = number of independent variables

The term linear regression stems directly from the use of the linear regression model presented here. One form of the equation, if plotted on a graph, would constitute points in a straight line.^{1/} The terms simple regression and multiple regression refer to the number of independent variables used in the analysis. A simple regression analysis relates only 1 independent variable to the dependent variable. Multiple regression analysis relates 2 or more independent variables to the dependent variable.

^{1/} There are a multitude of non-linear models used in non-linear regression analysis but these methods are beyond the scope of this publication.

Simple Linear Regression

The regression model for simple linear regression would be represented as:

$$Y = B_0 + B_1 X_1$$

Where: Y = dependent variable (from the sample)
B₀ = intercept (solved for by SAS)
B₁ = unknown parameter (solved for by SAS)
X₁ = independent variable (from the sample)

Assume there is interest in whether or not the age of the farmers sampled in Cook and Haynes Counties affects their use of conservation tillage. Also, if there is some effect on usage, how much is it affected? And finally, can the use of conservation tillage by similar farmers be predicted based on this information?

The SAS procedure "Reg" can be used in conjunction with the initial program presented in Figure 1, just as all SAS procedures have been used in previous chapters. By inserting "PROC REG;" and "MODEL ACRESMIN = AGE;" into lines 36 and 37, respectively, the program will generate output similar to that presented at the top of Figure 8. The program has calculated B₀ as 746.92 and B₁ as -11.11. Thus the regression model is:

$$\text{ACRESMIN} = 746.92 - 11.11(\text{AGE})$$

The negative B₁ implies that the older farmers in the sample conservation till less than the younger farmers. That is, observing the youngest to the oldest members of the sample, there is a downward trend in the use of conservation tillage. It is very important, however, to test the significance of both B₀ and B₁ to determine the degree of confidence in the results. The "Test of Significance" operates much like the Test of Difference and Testing the Mean. The REG procedure supplies the calculated t and degrees of freedom for the hypothesis that B₀ = 0 and B₁ = 0. If the hypothesis that B₀ = 0 is accepted, then the model becomes:

$$\text{ACRESMIN} = -11.11(\text{AGE})$$

If the hypothesis that B₁ = 0 is accepted, the equation falls apart and it must be assumed that the age of a farmer has no effect on his use of conservation tillage.

For this example, assume 99% confidence in the equation. With 23 degrees of freedom and a 99% confidence requirement an interpolation is necessary to find table t equal to 2.81, Table 2. The absolute value for the calculated t of both B₀ and B₁ (12.22 and 8.95) is greater than 2.81 so both hypothesis that the parameters are equal to zero are rejected. The equation is tested and remains:

$$\text{ACRESMIN} = 746.92 - 11.11(\text{AGE})$$

Figure 8. Relevant output from the Reg procedure, simple regression.
Records used: PROC REG; MODEL ACRESMIN = AGE;
MODEL ACRESMIN = EDUC;

Model: ACRESMIN = AGE

<u>Variable</u>	<u>Parameter Estimate</u>	<u>t for Hypothesis Parameter = 0</u>	<u>Degrees of Freedom</u>	<u>R-Square</u>
Intercept (B ₀)	746.92	12.22	23	.78
AGE (B ₁)	-11.11	-8.95		

Model: ACRESMIN = EDUC

<u>Variable</u>	<u>Parameter Estimate</u>	<u>t for Hypothesis Parameter = 0</u>	<u>Degrees of Freedom</u>	<u>R-Square</u>
Intercept (B ₀)	-319.86	-3.76	23	.65
EDUC (B ₁)	42.20	6.55		

Thus, the relationship described previously between AGE and ACRES MIN also stands. What does this equation mean? Given the age of a farmer with similar characteristics as those sampled, the equation can predict the amount of conservation tillage he is practicing. Thus, a 48 year old farmer could be expected to conservation till 214 acres, ($214 = 746.92 - 11.11(48)$). A 24 year old farmer could be expected to conservation till 480 acres, ($480 = 746.92 - 11.11(24)$)

For a second example, assume there is interest in whether or not the education level of farmers affect their use of conservation tillage. The bottom of Figure 8 is the result of using the initial program, Figure 1, with "PROC REG;" and "MODEL ACRESMIN = EDUC;" inserted in lines 36 and 37, respectively. The regression model becomes:

$$\text{ACRES MIN} = -319.86 + 42.20(\text{EDUC})$$

The absolute value of calculated t's for B_0 and B_1 (3.76 and 6.55) are each larger than the table t (2.81). Thus, the original equation remains since the hypotheses that $B_0 = 0$ and $B_1 = 0$ are rejected. A farmer with similar characteristics as those sampled with a high school education would be predicted to conservation till 187 acres, ($187 = -319.86 + 42.20(12)$).

The R-square is the estimated coefficient of determination. It is the fraction of total variation in the dependent variable that is accounted for by the independent variable. In the two examples, both age and education were separately tested significant (using the test of significance) in explaining the variation of acres conservation tilled. Age did a better job because the R-square of AGE is higher than that for EDUC, Figure 8. Why? The closer the R-square is to 1, the better job the model does of explaining the dependent variable, thus the more useful it is in prediction. (Studying the R-Square formula in the Appendix of Formulas will aid understanding the implications of the actual measure itself).

Multiple Linear Regression

The regression model for multiple linear regression is represented as:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$$

Where: Y = dependent variable (from the sample)
 B_0 = intercept (solved for by SAS)
 $B_1 - B_n$ = unknown parameters (solved for by SAS)
 $X_1 - X_n$ = independent variables (from the sample)

Using the conservation tillage example, the effects of both age and education on conservation tillage use could be found by substituting these variables into the regression model as:

$$\text{ACRESMIN} = B_0 + B_1 (\text{AGE}) + B_2 (\text{EDUC})$$

SAS solves for the parameters (B's) using the REG procedure with the insertion of the statements "PROC REG"; and "MODEL = AGE EDUC;" into lines 36 and 37 of the initial program, Figure 1, SAS outputs information similar to that in Figure 9. Using the calculated parameters presented in Figure 9, the model becomes:

$$\text{ACRESMIN} = 400.23 - 8.12(\text{AGE}) + 15.97(\text{EDUC})$$

Before this model is used for prediction, it is imperative to test the significance of each of the parameters (B's). At a 90% confidence level, for example, the table t is 1.72. Figure 9 illustrates calculated t's for all 3 parameters (absolute values) which exceed the table t. These values lead to rejecting the hypotheses^{1/} that the parameters equal zero, leaving the model as originally specified. -/

This model can be used to predict conservation tillage of a farmer with other characteristics similar to those in Cook or Haynes Counties. For example, a 48 year old farmer with a high school education would conservation till 202 acres according to this model, $(400.23 - 8.12(48) + 15.97(12))$. A 24 year old farmer with a 4-year college education is predicted to conservation till 461 acres, $(400.23 - 8.12(24) + 15.97(16))$. The predictive computation of a multiple linear regression model is much like that of a simple linear regression model, except that more than 1 independent variable is involved.

It is obvious, from Table 3, that the choice of independent variables is important to predicting acres conservation tilled:

Table 3. Predictions of conservation tillage for a 48 year old farmer with a high school education.

<u>Model</u>	<u>Prediction</u>
$\text{ACRESMIN} = B_0 + B_1(\text{AGE})$	214 Acres
$\text{ACRESMIN} = B_0 + B_1(\text{EDUC})$	187 Acres
$\text{ACRESMIN} = B_0 + B_1(\text{AGE}) + B_2(\text{EDUC})$	202 Acres

^{1/} If the hypothesis that B_2 (the parameter for EDUC) equal zero had been accepted, EDUC would be dropped from the analysis and a simple regression with AGE would be rerun, i.e. $\text{ACRESMIN} = B_0 + B_1(\text{AGE})$.

Figure 9. Relevant output from the REG procedure, multiple regression.
Records used: PROC REG; MODEL ACRESMIN = AGE EDUC;

Model: ACRESMIN = AGE EDUC

<u>Variable</u>	<u>Parameter Estimate</u>	<u>t for hypothesis Parameter = 0</u>	<u>Degrees of Freedom</u>	<u>R-Square</u>
Intercept	400.23	22	22	.81
AGE	-8.12	-4.38		
EDUC	15.97	2.08		

The test of significance will help to delete independent variables which do not add to the prediction, but how does one choose which independent variables to put into the model initially? The only person that can safely say which independent variables should be tested together is someone who is knowledgeable about the process being modeled. In this example, an expert in conservation tillage would be very helpful in listing the major variables which affect a farmer's decision to conservation till. In this example only 2 variables, age and education were considered. It is possible that a farmer's financial situation or dominant soil type could be as important. Therefore data would need to be gathered to include this information in the analysis and formulation of a predictive model.

To summarize, the basic steps to follow in studying the relationship between 2 or more variables in linear regression analysis are:

1. Consult an expert in the area being analyzed so that the major variables involved can be included in the sample and thus the model.
2. Using correct sampling techniques, collect the relevant data on each variable.
3. Construct a linear regression model in the general form $Y = B_0 + B_1 X_1 + B_2 X_2 + \dots + B_n X_n$. If only 1 independent variable is used, the model is one of simple linear regression. If more than one independent variable is included, the model is called a multiple regression model.
4. Use a SAS program similar to that in Figure 1 (with the appropriate variables and REG procedural cards) to calculate the parameters (B's).
5. Use the test of significance (with the chosen level of confidence) to test the desirability of each independent variable in the equation. If the hypothesis that $B_i = 0$ is accepted, then the independent variable should be dropped from the equation and the remaining model rerun and retested for B's = 0.
6. If the model is to be used for prediction, make sure that it is applied to populations with characteristics similar to those sampled. For example, using the model developed from a sample in Cook and Haynes Counties to predict conservation tillage in Mexico would be inappropriate because the characteristics of farmers in the U.S. vs. Mexico are so different. When using a regression model for prediction, keep in mind the characteristics of the population from which the sample and thus the model were derived.

VIII. APPENDIX OF FORMULAS

BASIC STATISTICS

Mean: $\bar{X} = \frac{\sum X_i}{n}$

where:

X_i = the observed value of the i^{th} unit in the sample

n = number of units in the sample

Variance: $V = \frac{1}{n-1} (\sum X_i^2 - \frac{(\sum X_i)^2}{n})$

where:

X_i = the observed value of the i^{th} unit in the sample

n = number of units in the sample

Standard Deviation: $S = \sqrt{V}$

where:

V = Variance

Coefficient of Variation: $CV = \frac{S}{\bar{X}} \cdot 100$

where:

S = Standard Deviation

\bar{X} = Mean

Standard Error of the Mean: $S_{\bar{X}} = \sqrt{\frac{V}{n} (1 - \frac{n}{N})}$

where:

V = Variance of the sample

n = Sample size

N = Population size

$1 - \frac{n}{N}$ = finite population correction

Correlation Coefficient: $r = \frac{\sum X_i Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}}$

where:

X_i = the observed value of the i^{th} unit in the X sample

Y_i = the observed value of the i^{th} unit in the Y sample

SAMPLING

Sample Size: $n = \frac{t^2 V}{E^2}$

where:

t = t value from students t distribution, Table 2

V = variance of the sample

E = acceptable error from the mean value

Sample Variance Estimate: $V = \left(\frac{R}{4}\right)^2$

where:

R = range of data expected

CONFIDENCE INTERVALS

Confidence limits: $\bar{X} \pm (t) (S_{\bar{X}})$

where:

\bar{X} = mean

t = t value, Table 2

$S_{\bar{X}}$ = standard error of the mean

HYPOTHESIS TESTING

Calculated t: $\text{calculated } t = \frac{|\bar{X} - H\bar{X}|}{S_{\bar{X}}}$
(Test of the Mean)

where:

\bar{X} = mean

$H\bar{X}$ = hypothesized mean

$S_{\bar{X}}$ = standard error of the mean

Calculated t:
$$\text{calculated } t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{PV (n_1 + n_2)}{(n_1)(n_2)}}}$$

(Test of Difference)

where:

\bar{X}_1 = mean of sample 1

\bar{X}_2 = mean of sample 2

n_1 = size of sample 1

n_2 = size of sample 2

PV = pooled variance; $PV = \frac{SS_1 + SS_2}{n_1 + n_2 - 2}$

where:

SS1 = sum of squares, sample 1

$$SS1 = \sum X_1^2 - \frac{(\sum X_1)^2}{n_1}$$

SS2 = sum of squares, sample 2

$$SS2 = \sum X_2^2 - \frac{(\sum X_2)^2}{n_2}$$

REGRESSION ANALYSIS

Beta parameters (B_1): $B_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}$

where:

$$x_{1i} = X_{1i} - \bar{X}_1$$

$$y_i = Y_i - \bar{Y}$$

Intercept (B_0): $B_0 = \bar{Y} - B_1 \bar{X}_1$

where:

B_1 = Beta parameter for variable X_1

Coefficient of Determination: $r^2 = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)}$

where:

$$x = X - \bar{X}$$

$$y = Y - \bar{Y}$$

Calculated t: $\text{calculated } t = \frac{(B_i - 0)}{S_{B_i}}$
(Test of Significance)

where:

B_i = beta parameter for X_i

S_{B_i} = standard deviation of B_i

$$S_{B_i} = S_E^2 / \sum (X_{1i} - \bar{X}_1)^2$$

where:

$$S_E^2 = \sum (Y - \hat{Y}_i)^2 / (n-2)$$

where:

\hat{Y}_i = value of estimated regression equation at the i^{th} value of X_1

IX. GLOSSARY

- calculated t - a confidence coefficient calculated from sample measures including the sample mean and standard error of the sample mean
- coefficient of variation (CV) - a relative measure of variation used to compare the variability of data sets that are in different units
- confidence limits - an interval estimate of a population measure e.g., confidence limits can express the true population mean as an interval estimate with a certain probability
- correlation - a qualitative correspondence between two sets of data either in a positive (move in the same direction) or negative (move in the opposite direction) manner
- correlation coefficient - a calculated coefficient ranging from -1 to 1 which measures the directional association between two sets of data
- data - compiled information which can be used for analysis or computation
- dependent variable - the variable in a linear regression model that is explained by one or more independent variables and appears on the left side of the equation
- hypothesis - an assumption subject to verification or proof
- hypothesis testing - using statistical procedures to accept or reject an initial hypothesis
- independent variable - the variable or variables in a linear regression model that explain the dependent variable and appear on the right side of the equation
- intercept - the B_0 parameter in a linear regression model which would appear as the intercept of a line on a two dimensional graph
- job control language (JCL) - the records of a computer program that allow access to the system; JCL records usually appear at the beginning of a program with an initial symbol of double slashes (//)
- linear regression analysis - the estimation of a relationship between two or more variables in a linear fashion

mean - an arithmetic average, usually used to describe the average value of a particular sample

multiple linear regression - estimation of a relationship between more than two variables (one dependent and two or more independent)

population - the whole; the entire set of items or individuals from which a sample is drawn

predictive model - the equation that results from the regression procedure of SAS; called predictive because new values for the independent variables can be used to estimate or predict a new value for a dependent variable

presampling - a mini sample used to roughly estimate the variance of a given population so that sample size can be calculated and applied to the principal sample when it takes place

PROC CORR - a SAS procedural record that initiates a program to calculate correlation coefficients for variables in the data set

procedural records - the portion of a SAS computer program that contains records calling for execution of prewritten SAS procedures

PROC MEANS - a SAS procedural record that initiates a program to calculate many of the basic statistical measures for a given data set

PROC PLOT - a SAS procedural record that initiates a program to plot designated variables on a two dimensional graph

PROC PRINT - a SAS procedural record that initiates a program to print out the data set

PROC REG - a SAS procedural record that initiates a program to estimate the regression parameters (B's) in linear regression analysis

PROC TTEST - a SAS procedural record that initiates a program to perform the test of difference

random sample - a sample chosen so that each value in the population has an equal and independent chance of being collected

regression parameters - known as beta coefficients (B's) in linear regression and solved for by SAS using the PROC REG procedure

sample - a portion of the whole regarded as representative of the whole; a collection of data used to represent the population

simple linear regression - estimation of a relationship between two variables (one dependent and one independent)

standard deviation - dispersion of values about the mean, indicating the variation in the data set

standard error of the mean - a measure that indicates the variability of sample means much like the standard deviation indicates variability in individual sample observations

Statistical Analysis System (SAS) - a computer software system originally developed for statistical needs; it contains a collection of prewritten programs that can be accessed by a few simple commands

statistics - the science that deals with techniques used to obtain analytical measures, the methods for estimating their reliability, and the drawing of inferences from them

t or table t - a confidence constant usually found in a "t" table and originally developed from a probability distribution called "students t"; t is used in many statistical calculations including sample size, confidence limits, hypothesis testing, and regression analysis

table of random numbers - table with a large quantity of single digits arranged in a random fashion; used to facilitate the design of a random sample

test of difference - a test that checks for differences in two populations by statistically comparing sample means

test of the mean - a test that uses the mean from a sample and probability theory to accept or reject a hypothesis about the sample

test of significance - a test of the desirability of each independent variable in regression analysis based on the hypothesis that the beta parameters (B's) equal zero

variable - a characteristic of a population that can be chosen for study

variance - the square of the standard deviation used to measure data variability

X. REFERENCES

Anderson, Richard L.; Allen, David M. and Cady, Foster B., 1972. "Selection of predictor variables in linear multiple regression" In Statistical Papers in Honor of George W. Snedecor. Ames, Iowa: The Iowa State University Press.

Bancroft, T. A., 1968.: Topics in intermediate statistical methods. Ames, Iowa: The Iowa State University Press.

Larson, Harold J., 1974. Introduction to probability theory and statistical inference. (New York: John Wiley and Sons, 1974).

McConnell, Campbell R., 1981. Economics, 8th ed. (New York: McGraw-Hill Book Company, 1981).

Ostle, Bernard and Mensing, Richard W., 1975. Statistics in research, 3rd ed. Ames, Iowa: The Iowa State University Press.

Snedecor, George W., 1980. Statistical methods, 8th ed. Ames, Iowa: The Iowa State College Press.

SAS Institute Inc., 1982. SAS user's guide: Basics, 1982 Edition, Cary, NC: SAS Institute Inc., 923 pp.

SAS Institute Inc. 1982. SAS user's guide: Statistics, 1982 Edition. Cary, NC: SAS Institute Inc., 584 pp.