Part 610 – Natural Resource Economics Handbook

Subpart F – Introductory Statistics

610.50 Introduction

A. What is Statistics?

(1) Statistics is "The science of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling." Measured population characteristics, such as the population mean, are called "parameters." A measured characteristic of a sample, such as the sample mean, is called a "statistic."

(2) The science of statistics deals with—
(i) Designing surveys.
(ii) Collecting and summarizing data.
(iii) Measuring the variations in survey data.
(iv) Measuring the accuracy of estimates.
(v) Testing hypotheses about the population .
(vi) Studying relationships among two or more variables.

(3) Statistics, then, is the science of providing the techniques used to obtain analytical measures, the methods for estimating their reliability, and the drawing of inferences from them.

B. Purpose of This Subpart

The purpose of this subpart is to facilitate the understanding of statistical concepts and techniques, their limitations, the assumptions behind them, and the interpretations that can be made from them. It is a short reference guide for those who apply and use statistical analysis in their work. Users of this subpart are, however, still encouraged to consult experts when there is a need for data collection, analysis, and interpretation.

C. Uses for Statistics in NRCS

NRCS employees are responsible for answering questions and solving problems related to conservation work for land users and resource managers. The keys to meeting these needs are to acquire relevant information and data and to properly interpret the collected information. Hence, statistics—when defined as a tool for data collection, analysis, and interpretation as well as prediction—is useful for NRCS employees in general and particularly for those technical experts who are responsible for the use and analysis of resource information.

D. Data Collection (Sampling)

(1) NRCS personnel are constantly weighing whether or not additional information (obtained at some cost) will yield results that justify the expense of collecting it. Some problems encountered are too minor to justify a large expenditure for detailed data collection. On the other hand, some decisions involve tens of thousands or millions of dollars and years of development. It's only common sense to collect useful information to help make those types of decisions. But where is the line drawn? How far should one go to ensure the accuracy of data collection efforts?

(2) In most situations, it is impossible to collect information from every data source. For example, NRCS may need to find out the extent to which conservation tillage is

currently being applied in the United States. It would be almost impossible to contact and interview every farm operator without committing huge resources. 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, taking a larger sample than necessary could waste funds. Statistical methods can be used to help answer these questions. With the aid of statistics, the sample size that meets specific precision criteria for individual data collection can be determined.

E. Data Analysis

(1) When data are collected from a sample, whether from a primary or secondary source, confidence in the data is always a question. It is possible, through the use of statistical methods, to quantify "confidence" in the reliability of the sample data characterizing the data in a population. In fact, many of the basic statistical measures can be tested for their degree of reliability, a unique characteristic of statistical methods.

(2) Why is the ability to quantify confidence in statistical measurements so important? Imagine assigning two people to independently study a particular problem. To make an intelligent decision, both people separately study the problem, collect and analyze the data presumed to be relevant to studying the problem, and report their results. Unfortunately, their results are very different. Person 1 subjectively decided to take a 10-percent sample from the targeted population. He averaged the data and reported the averaged results, or means. Person 2 used a statistical method such as sampling to determine the most economical sample size out of the total. His averages (means) of sample characteristics were reported as well, but he also included the limits of the statistical confidence that his sample means were representative of the actual mean of the population. Because the second person quantified the limits of confidence in his results, indicating levels of reliability, his report was more trustworthy than the first person's report.

F. Interpretation and Prediction for Decisionmaking

(1) Most NRCS data are eventually interpreted to help make present decisions or to predict future trends of events for future decisions. Hypothesis testing of statistics is a method that can be used to compare an estimate or intelligent guess to actual sampled data. For example, if 33 bushels per acre seems like a low crop yield, a quick sample could be developed from the Agricultural Census data or some other obtainable sources in order to test whether that is actually the case. 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.

(2) Most NRCS technical specialists are expected to perform predictions for the future when comparing situations of "with" and "without" the application of conservation measures. A statistical analysis technique called regression analysis enables users to predict future trends based on past data. For example, it is possible to estimate future yields based on past yields and inputs in a farm. The relationship between yields and associated inputs is commonly known as a production function, in which the predicted yields are dependent on the expected farm inputs in a mathematical equation or function. This is one of the most powerful tools in statistics, and it is probably the most widely used.

(3) To summarize, statistics gives the user—

(i) A number of ways to sample and analyze data.

(ii) Relatively simple, timesaving means to quantify the reliability of data derived from the sample.

(iii) Techniques to compare estimates and data that represent those of the population.
(iv) A way to predict future trends based on past events.

610.51 Elements of the Sampling Process

A. The process of using statistics sampling to answer questions and make decisions involves a number of steps:

(1) Define the problem
(2) Design the sampling procedure
(3) Collect and analyze the data
(4) Interpret the results

B. Define the Problem

(1) Statistical sampling and analysis is a process that begins with a problem or set of questions. What are the problems or questions you are trying to answer? What types of data do you need to measure or collect to answer those questions? What is the area of interest: a watershed, a set of counties, or an entire State? Answers to these questions will help to clearly define the problem.
(2) The problem must be specified in a way that can be answered through statistical processes. The population and measurements needed to answer the question will be defined at this stage. The population, the set of all elements about which we wish to make an inference, must be clearly defined before the data are collected. For example, if you wish to estimate the average size of farms in a three-county area, the population would be the farms in those three counties. You must decide if farms will be defined in terms of ownership units, management units, or some other method. Once the definition is determined, it must be consistently adhered to throughout the sampling process. Similarly, careful thought and definition must be given to the characteristics measured for each sampling unit.
(3) It may not be necessary to conduct your own survey. Contact experts in your State office for assistance in learning what data are available for your area. There are several national surveys whose data or summary tables may be utilized.
(i) The National Resources Inventory (NRI) is conducted by the NRCS on non-Federal land in the United States. The NRI summary tables may provide the information needed, especially related to land use area and erosion. Your State NRI specialist will be able to assist you with specific summaries.
(ii) The National Agricultural Statistics Service (NASS) Census of Agriculture provides a sample of more than 70 percent of U.S. farms, including resource, production, demographic, and economic information. The other surveys conducted by NASS are useful information about agriculture prices, livestock and crop production, and chemical use.
(iii) The U.S. Census Bureau database provides access to population and demographic data dating back as far as 1790.
(iv) The Forest Inventory and Analysis (FIA) is conducted by the U.S. Forest Service and provides detailed information on forestland.

C. Design the Sampling Procedure

(1) One way to determine the average size of farms in a State would be to find the acreage of each farm in the State and calculate the average. Conducting a census is a full data collection effort using a set of measurements for an entire population. Although a census provides complete information on the set of measurements taken

for a population, it is often not practical because of the time and money needed to conduct such a survey. Statistical analysis enables a researcher to learn a great deal about a population without having to resort to committing the time and money required to gather data for the entire population.

(2) Sampling may be used to collect information on a subset of the population. It is important that the sample be representative of the population. If we drew our sample from a list of corporate farms, we would be missing representation from many small farms. Similarly, if we drew a sample from one area of the State, the sample would not be representative of farms across the entire State. There are a number of methods used to select a representative sample; these methods are called "sample survey designs" or "sampling designs." One of these, simple random sampling, is discussed in this subpart (see note below). Many other sampling designs exist that allow the selection of a representative sample that may be used to make an inference about the entire population. These designs can be explored further by consulting statistical literature and experts.

(3) The data should be collected in a way that follows the sampling design, should be organized for ease of analysis, and should be checked for recording errors. During this stage, detailed data collection and storage tools (e.g., worksheets or spreadsheets) must be acquired or developed. Statistical methods for analysis of data that are appropriate for the sample design should be determined as well. Ideally, a small pilot test of the entire process will be conducted at this stage to find any flaws in the process before it is fully implemented.

Note: Note that the computing equations for the statistics and example calculations using Excel presented in this subpart are only appropriate for summarizing data collected with a simple random sample.

D. Collect and Analyze the Data

Implement the detailed plans made during the design stage. Careful planning during the design stage will help data collection and analysis to run smoothly. However, data should be checked periodically during collection to discover any problems, such as incorrectly calibrated instruments, early in the process.

E. Interpret the Results

The sampling process began with a problem and a set of questions that needed to be answered. The results of the statistical analysis must be reported in a way that will answer those questions and provide all pertinent details. If the yields of two brands of corn are compared, report not only the brands, but also the number of fields on which the tests were conducted, the geographic range of these fields, etc. In presentation of statistics, always accompany any summary statistic with its appropriate measure of reliability. For example, if you report the average yield of corn, include the standard error of the mean yield, or, alternatively, include a confidence interval showing the level of confidence (e.g., 90 percent, 95 percent, or 99 percent).

610.52 Sampling Techniques

A. 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 in market analysis, consumers are surveyed to discover their wants. A carload of coal or grain is accepted or rejected based on analysis of a few pounds. Physicians make decisions about health based on records obtained from a few hospitals. The use of sampling is widespread,

yet the importance of sampling sometimes goes unnoticed. In NRCS, day-to-day duties are often fulfilled with studies of raw data sets, such as the NRI and soil surveys, that originated as a survey or sample of some kind.

B. A sample is basically a small collection of information from some larger aggregated population. The sample is collected and analyzed to make inferences regarding the relevant characteristics of the total population. One thing that makes this process difficult is the presence of variation. If all farmers on Earth were alike, a sample of one farmer would represent all farmers. Since this is not the case, members of the population are usually different and, therefore, successive samples are usually different. Thus, the major challenge of statistics is to reach appropriate conclusions about the population in spite of sampling variation.

C. The ultimate goal of sampling is to make an inference about the population as a whole without measuring all of the elements belonging to it. There are a wide variety of sampling designs available that will enable us to collect a sample of the population that will be adequately representative. One of these techniques, simple random sampling, is presented in this subpart because it is easy to understand and apply. However, we should be aware that other sampling designs might allow you to collect data on a smaller number of observations and still obtain the same level of information that you would with a larger number of observations using simple random sampling. Work with a statistician to help you design the most efficient sample for your specific situation. A well-designed survey can save both time and money and produce estimates that are reliable.

D. Simple Random Sampling

(1) Simple random sampling is designed in such a way that every element in the population has an equal chance of being sampled. Careful planning ensures that this criterion is met. If certain elements of the population have no chance of entering the sample, we cannot say anything about those elements with any certainty. For example, suppose we want to conduct a study to determine the favorite TV show of residents in a certain city. If we only went to the middle schools and interviewed all the students in these schools, would we get a fair representation of residents in the city? Suppose we want to conduct a study to estimate the proportion of cultivated cropland that used conservation tillage in a given year. If we collected our sample from fields adjacent to roads, can we say anything about tillage in the nonadjacent fields?

(2) It may be necessary to begin with a list of the population. For example, field offices often need to sample tracts of land enrolled in various conservation programs within a county to check for compliance. Simple random sampling may be used to draw a sample from a list of signup sheets. Then a random number generator, such as the one in Excel, may be used to select a sample from that list. In this way, each enrolled tract in the county would have an equal chance of entering the sample. Often a population cannot be enumerated by a list. The following example describes simple random sampling of a geographic area.

E. Earthworms: An Example of Simple Random Sampling

(1) Suppose you want to estimate the average number of earthworms per square foot in the top 6-inch soil layer of a 10-foot by 20-foot garden. You set up a grid of lines at 1-foot intervals to get a total of 200 square foot plots. Sketch the layout on a piece of paper and number the plots from 1 to 200.

(2) Use the Excel random number generator to select a set of 10 plots to be sampled. Type the following command in a cell: =RANDBETWEEN (1,200). Copy and paste

this command into the nine cells below the first. If any of the values are repeated, you may use the command to select a new value to replace the redundant value. This is called sampling with replacement.

(3) Once you have the set of 10 random numbers, mark these on the paper sample grid. Find the corresponding locations in the garden and mark them. Record the number of earthworms in the top 6 inches of soil for the selected plots. In the random sample of 10 square-foot plots (6 inches deep) within a garden, the number of earthworms counted in each plot was: 2, 5, 6, 6, 7, 9, 10, 11, 12, and 15. We will use this example in calculating some of the statistics later in this subpart.

F. Sample Size

(1) Before using the random number table or Excel to select a sample, one must decide how large a sample will be needed. Too small a sample may lead to inaccurate observations 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 would require undue expense and time.

(2) How to determine a sample size? The sample size is a function of confidence constant t, the variance of sample mean, V, and the acceptable error, e, from the true mean as follows:

n= size of the sample = tV/e2

(3) Step 1: Specify the level of confidence required in the sample mean.
(i) t = confidence constant: the level of confidence required in the sample mean
(ii) t = 1σ provides 67-percent confidence level
(iii) t = 2σ = 4 for 95-percent confidence
(iv) t = 3σ = 6.76 for 99-percent confidence level

(4) Step 2: Specify the acceptable error, e, from the true mean of the population, or the variance of the sample mean, V.
(i) e = acceptable error from the true mean
(ii) V = variance = (R/t)2
(iii) R = the range of data expected

(5) 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 percent 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 for the sample.

(6) The variance of the sample V is calculated from the sample data. So, how can an estimate of V be used to calculate sample size if no data have been collected yet? The best way to obtain V is to take a presample, calculate that variance, and use it in finding the sample size. But presampling is normally costly and time consuming. Thus, the following example was developed to show how to find a rough estimate of V without presampling.

(7) In the conservation tillage example, the population consists of 700 farms that are each 600 acres in size. The main variable of interest is the acreage that is conservation tilled. It's quite obvious that this amount could range from 0 to 600, because a farmer could either conservation till all or none of his or her acres or some number in between. Using the V formula with R= 600-0=600, V= (600/4)2 = 22,500:

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

n = = tV/e2 = (4)(22,500)/(50)2 = 36

(9) To demand a 99-percent degree of confidence, and to require estimates to be within 40 acres of the true mean, would necessitate a sample of 95 farmers rather than 36:

n = (6.76) (22,500)/(40)2 = 95

(10) Generally, as the estimated variance and required degree of confidence increase and the acceptable error decreases, sample size must increase.

610.53 Basic Statistical Concepts

A. Bias, Accuracy, and Precision

(1) The terms "bias," "accuracy," and "precision" are commonly used—and often misused—when discussing sampling estimates. Bias is a systematic distortion that may be caused by a poorly constructed sampling design or a flaw in measurement. Suppose we want to estimate the average pH of soil in a field. An average value calculated from soil pH data obtained using an improperly calibrated hand-held pH meter would be biased.

(2) Accuracy is a measure of success in estimating the true value of a quantity. We rarely know what the true value of a quantity is. Sound statistical practices and careful measurements will help us get closer to the truth.

(3) Precision is the spread of the sample data about their average value. In our example, a field containing both areas of highly acidic soil and areas of alkaline soil would likely produce a wide range of pH values. The precision of the pH readings from this field would be lower than that of a field with more uniform pH readings. In summary, a seriously biased estimate of soil pH may be precise, but it cannot be accurate.

B. Types of Data

(1) Data are recorded for characteristics of interest for each sampling unit that is included in the sample. A characteristic that varies from one sampling unit to another is referred to as a variable. The type of data that are recorded for these variables will affect the type of analysis that may be conducted on the data. The types of data are ratio scale, interval scale, ordinal scale, and nominal scale.

(2) **Ratio scale** data have a constant unit interval and zero is a point on the scale used to measure these data. Examples of this type of data are number of acres, bushels per acre, height, and weight. Let's consider the number of acres in terms of the definition above. Each acre is the same size, and it is physically possible to have zero acres.

(3) **Interval scale** data have constant unit intervals, but not a true zero. Celsius or Fahrenheit temperature scales are an example of interval scales. Each Celsius or Fahrenheit degree is a constant interval unit, but zero is arbitrary. The Kelvin temperature scale on the other hand has a constant interval unit and a physically meaningful zero. The Kelvin temperature scale is a ratio scale.

(4) **Ordinal scale** data do not have a constant interval but may be used in relative comparisons. Some examples of these ordered data are height recorded as high, medium, or low; a preference rating on a scale of 1 to 10, where 1 is least preferred and 10 is most preferred; and brightness recorded as bright, brighter, or brightest.

(5) **Nominal scale** data are classified by name only with no relative order for comparison. Examples of nominal scale data include brand of seed corn, questionnaire responses of yes or no, and eye color (e.g., blue, green, and brown).

(6) **Conservation Tillage Sample Data Set**.—This data set contains hypothetical data from 25 farms, each 600 acres in size, from two counties. It contains data on observation number (OBS), acres conservation tilled (ACRES), age of operator (AGE), years of formal education (EDUC), county (CO), and "uses conservation tillage" (CONSTL). This data set is shown below as it would appear in an Excel spreadsheet and will be used in many subsequent examples in this subpart.

Table F-1 Conservation Tillage Sample Data Set

	A	B	C	D	E
1	OBS	ACRES	AGE	EDUC	COUNTY
2	1	190	50	14	Cook
3	2	135	59	14	Haynes
4	3	275	39	16	Haynes
5	4	185	49	14	Haynes
6	5	340	39	12	Cook
7	6	575	32	16	Haynes
8	7	210	51	14	Cook
9	8	95	55	8	Haynes
10	9	55	67	8	Cook
11	10	210	28	12	Haynes
12	11	280	35	14	Haynes
13	12	0	68	8	Cook
14	13	120	55	12	Haynes
15	14	80	59	12	Cook
16	15	600	30	18	Cook
17	16	415	29	18	Haynes
18	17	0	62	8	Haynes
19	18	395	27	16	Haynes
20	19	480	31	16	Cook
21	20	180	52	12	Cook
22	21	60	63	6	Haynes
23	22	0	61	12	Haynes
24	23	108	61	12	Haynes
25	24	225	46	12	Cook
26	25	295	37	16	Cook

C. Plot the Data

Plotting the data allows the analyst to get a mental picture of the characteristics of the data set and is an important tool to be used before, during, and after analysis. These plots will assist with visualizing the value around which most of the data are concentrated, the spread of the data, and any values that are much larger or smaller than the majority of the observations. Relationships between two variables may become evident through viewing of scatter plots. Several types of plots commonly used for this purpose are line plots, stem-and-leaf plots, histograms (bar charts), box plots, and scatter plots. Histograms and scatter plots can be created using Excel.

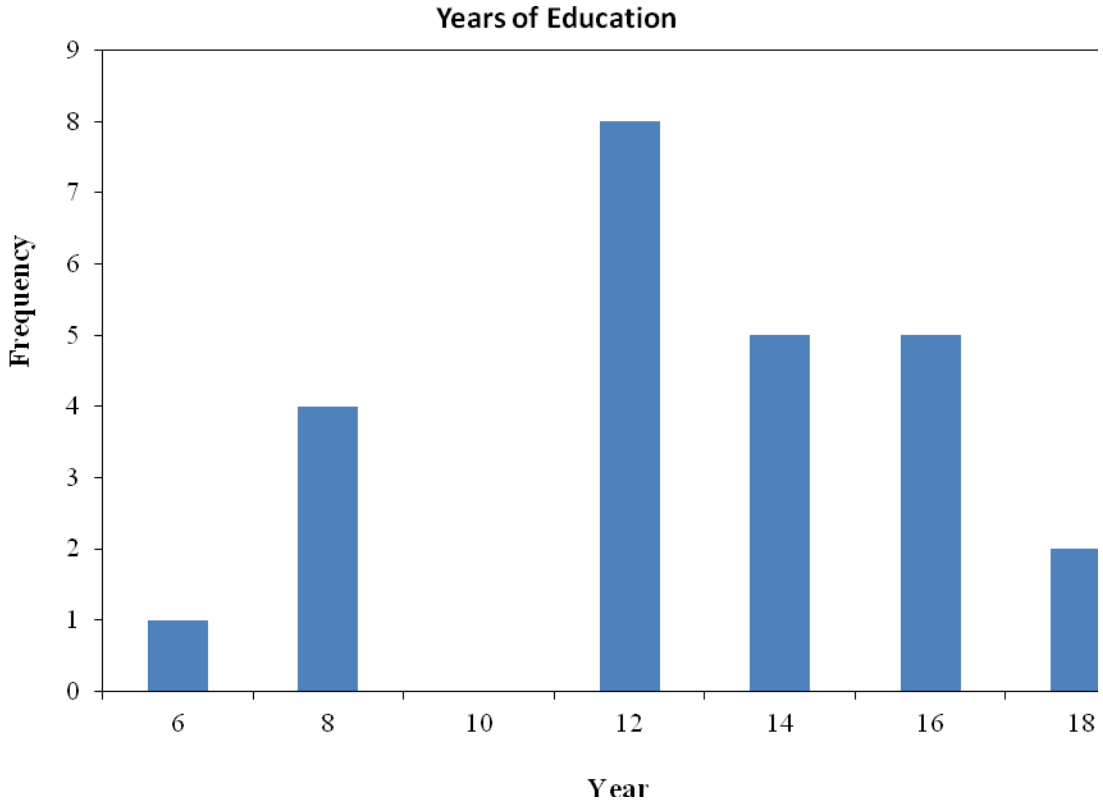
D. Histogram

(1) A histogram displays the frequency of values for observations in a dataset. It provides a picture of the distribution of data. Preparing a histogram enables us to

easily see the minimum, maximum, and most common values in a data set. We may be able to use a histogram to predict what the calculated value of the average will be.

(2) Figure 610F-1 is a histogram displaying the number of years of education for farmers included in the Conservation Tillage example. Two farmers had 18 years of education and one had 6 years. The range of data is 12 (18-6). Most farmers in this example (eight in number) had 12 years of education. From this histogram we can guess that the average farmer in this example has at least 1 year of post-high-school education.

Figure F-2 Histogram of Years of Education for Farmers in the Conservation Tillage Example



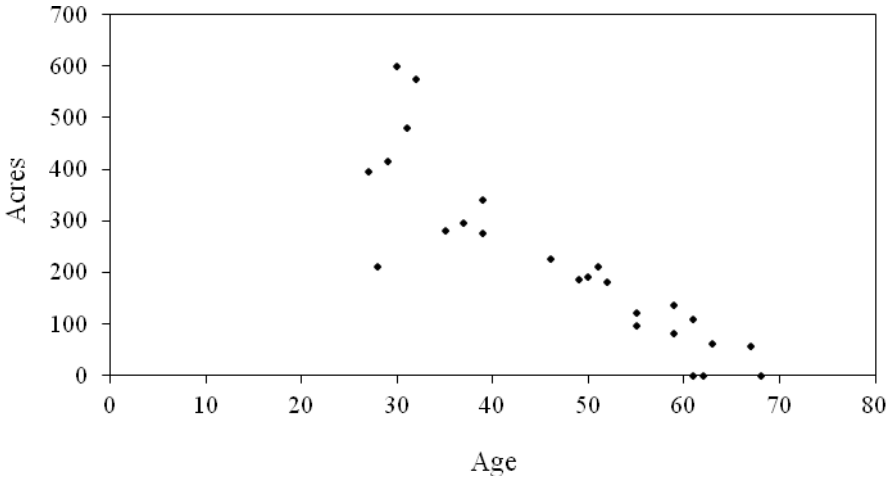
(3) To construct this chart in Excel, you need to use the histogram tool in the Excel Analysis ToolPak. If you don't see the "Data Analysis" button in the "Analysis" group on the "Data" tab, you must load the Analysis ToolPak add-in (see note below). Set up a column containing the range of education (6 to 18) with 2-year increments in cells F1-F7 (you may choose 1-year increments or some other equally spaced values). These are the bin values. Select "Histogram" from the choices in "Data Analysis" drop-down list. Provide the Input Range (C1:C26), Bin Range (F1:F7), and Output Range (H1). Check the boxes next to "Labels and Chart Output" and hit "OK."

Note: The data analysis tools are CCE-approved as part of Excel and Access. In case your office has not had the Excel Data Analysis Tools loaded as part of the CCE Office installation, you should follow the "Excel Help" instructions for loading it. All NRCS technical specialists should have it available.

E. Scatter Plots

(1) Scatter plots provide a picture of the relationship between two variables. Figure 610F-2 is based on the conservation tillage example. The plot of acres versus age exhibits an inverse relationship between the number of acres in conservation tillage and the age of the farmers. In this example it appears that older farmers tend to use conservation tillage on fewer acres than do younger farmers.

Figure F-3 Scatter Plot Of Acres Of Conservation Tillage Versus Age Of Farmers In The Conservation Tillage Example



(2) The scatter plot is constructed using the "Scatter" button in the "Charts" group on the "Insert" tab in Excel. To construct this chart in Excel, highlight all data for acres and age in cells B1-C26, press the "Scatter" button and select the type of scatter plot with no lines.

610.54 Basic Statistics and Computation for Simple Random Sampling

A. Introduction.—This section presents a number of basic statistics describing the central tendency and variability that are derived from simple random sampling from the targeted population. Sampling designs determine the methods appropriate for computing these statistics. Other sampling designs require the use of alternative computing equations and possibly other statistical software to correctly process the data. For example, NRI data are collected using a stratified cluster sample. Calculating the means and standard errors for these data using the techniques presented in this section will lead to erroneous results. Work with your local NRI specialist to correctly summarize NRI data. Similarly, if you obtain data from other sources, always ask what sampling design was used to collect the data and how to correctly summarize it. The equations presented in the subsequent sections are common to sampling literature including Cochran (1977) and Schaeffer et al. (1996).

B. Mean, Median, and Mode

(1) The three measures of typical central tendency values are arithmetic mean, median, and mode. Arithmetic mean is probably the most well-known and often used statistic. It is referred to as the sample mean when it is derived from a sample. The mean is merely the arithmetic average of all the values from a sample and is intended to represent the "typical" value of the drawn population. Its advantages are ease of

computation, common usage, and use in algebraic manipulation. But the major disadvantage of this method is that arithmetic mean is unduly affected by extreme values and may, in fact, be far from representative of the data in the population. It is important, then, to be able to diagnose the variability of the data as well as the mean. Samples from data populations such as real estate values or annual incomes often contain extreme values. The sample mean is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} \qquad \text{(F-1)}$$

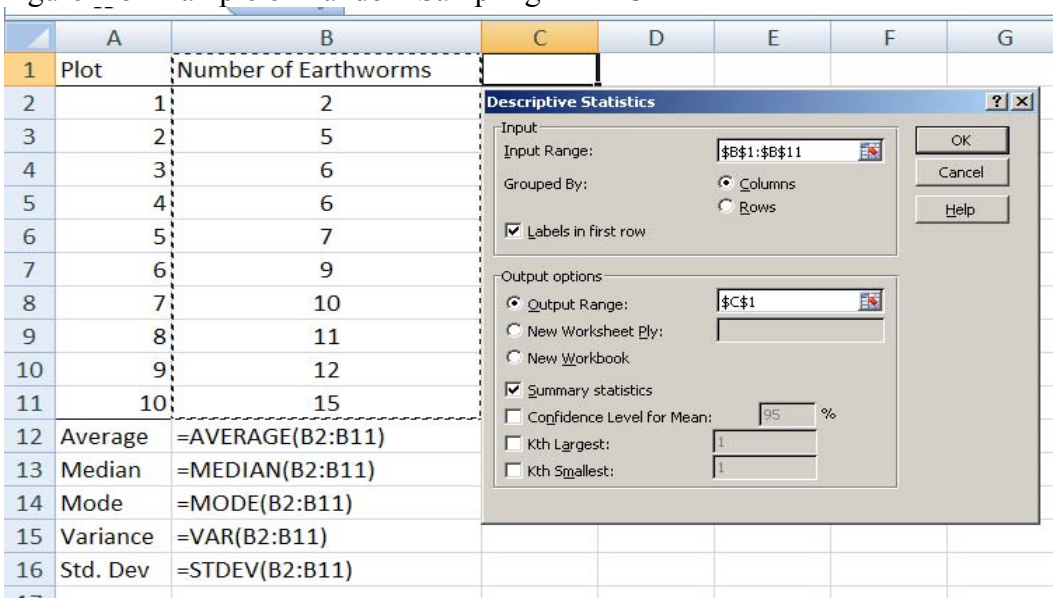
(2) In the random sample of 10 square-foot plots (6 inches deep) within a garden, the number of earthworms counted in each plot was: 2, 5, 6, 6, 7, 9, 10, 11, 12, and 15. The average number of earthworms per square foot in the top 6-inch layer of soil for the garden, $\bar{x}$, is 8.3. In Excel, place the 10 sample values in cells in a column (e.g., B2:B11) as in the figure below.

Figure F-4 A Random Sample Of Earthworms

	A	B
1	Plot	Number of Earthworms
2	1	2
3	2	5
4	3	6
5	4	6
6	5	7
7	6	9
8	7	10
9	8	11
10	9	12
11	10	15
12		

(3) Type the following command in cell B12: =AVERAGE (B2:B11). Alternatively, you may use the Data Analysis button in the "Analysis" group on the "Data" tab. Select "Descriptive Statistics" from the choice list. Enter the input range (e.g., B2:B11), output range (e.g., C1), check the boxes next to "Columns," "Labels in First Row," and "Summary Statistics," and hit "OK." The set of descriptive statistics calculated for the 10 selected numbers include mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, and count, as shown in the figure below.

Figure F-5 Example of Random Sampling in EXCEL



	A	B	C	D	E	F	G
1	Plot	Number of Earthworms					
2	1	2					
3	2	5					
4	3	6					
5	4	6					
6	5	7					
7	6	9					
8	7	10					
9	8	11					
10	9	12					
11	10	15					
12	Average	=AVERAGE(B2:B11)					
13	Median	=MEDIAN(B2:B11)					
14	Mode	=MODE(B2:B11)					
15	Variance	=VAR(B2:B11)					
16	Std. Dev	=STDEV(B2:B11)					

Figure F-6 Statistical Summary of a Random Sample

	A	B	C	D
1	Plot	Number of Earthworms	*Number of Earthworms*	
2	1	2		
3	2	5	Mean	8.3
4	3	6	Standard Error	1.21
5	4	6	Median	8
6	5	7	Mode	6
7	6	9	Standard Deviation	3.83
8	7	10	Sample Variance	14.68
9	8	11	Kurtosis	-0.27
10	9	12	Skewness	0.16
11	10	15	Range	13
12	Average	8.3	Minimum	2
13	Median	8	Maximum	15
14	Mode	6	Sum	83
15	Variance	14.68	Count	10.00
16	Std. Dev	3.83		

(4) The median value provides a second measure of central tendency. The median value of a data set is the midpoint or middle value of the dataset when it is ordered by size of those values. If the dataset has an even number of observations, the median value will be the average of the two middle values.

(5) In the example above, the number of earthworms counted in each plot was 12, 15, 9, 6, 7, 6, 10, 11, 2, and 5. To determine the median value of a set of numbers, the data must first be ordered by size as 2, 5, 6, 6, 7, 9, 10, 11, 12, and 15. This sample has an even number of observations. The median value, 8, is the average value of the middle two observations, 7 and 9. To calculate the median of the 10 values in Excel, type the following in cell B13: =MEDIAN(B2:B11).

(6) There are times when it may be better to use the median than the mean when reporting "typical" values. As a simple example, suppose the annual incomes (in thousands of dollars) for a random sample of 10 individuals working for a company are 23, 23, 24, 25, 32, 33, 33, 35, 36, and 150. The mean annual income of these 10 individuals is 41.4 thousand dollars and the median annual income is 32.5 thousand dollars. Which of these two measures might better represent the typical annual income of employees in that company? Why?

(7) Mode is the most frequent value observed in a dataset. In the earthworm count example, the mode is 6. To calculate the mode of the 10 values in Excel, type the following in cell B14: =MODE (B2:B11). The Excel formulas for mean, median, and mode are displayed in the above figures as well.

C. Sample Variance and Standard Deviation

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

(2) An advantage of these measures of dispersion is that they can be used to formulate confidence limits and hypothesis tests (to be discussed later). The major disadvantage becomes apparent when comparing two sets of data because the formula for a standard deviation dictates that the standard deviation measurement of a data set with relatively larger values will be higher than that of one with relatively smaller values. Consider the following examples: The mean of data set A is 15 and the mean of data set B is 1,000. The standard deviation is 9.78 for data set A and 20.09 for set B. It would be inaccurate to conclude that data set B is more variable than data set A. Therefore, a method of measurement is needed for comparing the variability of data sets with widely differing means.

(3) The standard deviation and variance of a data set 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 a function of the size of the mean. Therefore, it is meaningless to compare standard deviations between two different variables, say conservation tillage acres and farmers' age, as the magnitudes of the means are so different. It may be useful in some cases, such as when comparing the standard deviation of conservation-tilled acres in one sample to the standard deviation of conservation-tilled acres in some other sample. This is acceptable and would aid in choosing between two sets of data to study.

(4) The sample variance represents the variability in individual sample observations.

The sample variance, s^2, is used in calculating the estimated variance for the estimated population mean and total, the corresponding confidence intervals, and in calculating the sample size needed to estimate population means and totals. The sample variance for a simple random sample is calculated as (Cochran 1977, Schaeffer et al., 1996):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}. \qquad \text{(F-2)}$$

(5) The standard deviation for a simple random sample is the square root of the sample variance.

(6) The sample variance and standard deviation may be calculated in Excel, as shown above. In cell B15 type =VAR (B2:B11) to calculate the sample variance or =STDEV (B2:B11) to calculate the standard deviation. The sample variance for this example is 14.68 and the standard deviation is 3.83 earthworms per square foot of soil (6 inches deep).

D. Coefficient of Variation

(1) It may be useful in some cases to compare the variability of two different sets of data, such as comparing the total number of conservation-tilled acres in one sample to that of another sample. In this case, we would like to use a measure of relative variation—the coefficient of variation (CV)—to compare the variations of variables across different data sets. CV is a measure of relative variation in each set of data and is calculated by dividing the standard deviation by the mean, then multiplying 100. Using the example cited previously, the mean of data set A is 15 and the mean of data set B is 1,000. The standard deviation is 9.78 for data set A and 20.09 for data set B; therefore, 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 data set B, even though the standard deviation of A is less than that of B.

(2) 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.

(3) Overall, if one were 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 than the standard deviation.

E. Standard Error of the Mean

(1) The standard error, which can be thought of as the standard deviation of sample means, is a measure used extensively in the development of confidence intervals.

(2) The standard error of the mean is a measure that indicates the variability in sample means, much as the standard deviation indicates variability in individual sample observations. In the earthworm example, a simple random sample was taken from a population. If the random sample were to be repeated, the mean would probably not be 8.3. In fact, numerous samples from the same population would yield different means.

(3) It would be possible to estimate the variability of sample means by actually taking repeated samples and using the standard deviation formula. But the calculation of the standard error of the mean makes it possible to gain the same information by using the one initial sample. The standard error of the mean is calculated by dividing the standard deviation by the square root of the sample size. The standard deviation in the earthworm example is calculated as 3.83 earthworms per square foot of soil (6 inches deep). Dividing this result by the square root of 10 (the sample size) gives a standard error of 1.21 earthworms per square foot of soil (6 inches deep). Recall that the standard error is one of the statistics reported in the descriptive statistics found in the data analysis tools of Excel.

(4) Standard Error = $s_d = \sqrt{s/n}$ (F-3)

F. Correlation Coefficient

(1) The correlation coefficient is a statistic that measures a specific type of relationship between two variables. There are several different ways to measure the correlation between two variables. A commonly used correlation coefficient is Pearson's correlation coefficient. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. A sample-based Pearson's correlation coefficient for variables x and y is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{(F-4)}$$

(2) The correlation coefficient ranges numerically between 1 and -1.

(3) 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 levels might be collected. By looking at the relationship between data on farmers' ages and data on their education, a positive correlation coefficient close to one would suggest that older age is positively associated with higher education. A negative correlation coefficient close to -1 would indicate that older age is negatively associated with education. And finally, a correlation coefficient close to zero would indicate that the data do not support concluding that there is any significant correlation between a farmer's age and his or her education.

(4) 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 have higher income 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 income may buy more education just as they buy more automobiles and more vacations. Thus, higher income may cause higher education! Do not fall into the trap of extending correlation to causation. Causation will be discussed in the section on the use of regression analysis.

(5) Let's go through an example of calculating the correlation coefficient using data in the conservation tillage example. A scatter plot (figure 610F-2) displayed the negative relationship between these two variables. From that plot we know that the correlation coefficient between age and acres of conservation tillage will be negative. To calculate the correlation coefficient between age and acres, select the "Correlation Wizard" from the drop-down list of the "Data Analysis" button in the "Analysis" group on the "Data" tab, type B1:C26 in the "Input Range" box of the Correlation Wizard, check "Columns," "Labels in First Row," and "Output Range," type A28 or any other empty cell in the "Output Range" box, and click "OK." The resulting correlation coefficient between the two variables is -0.88.

610.55 Confidence Limits (Reliability of a Sample)

A. 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 and it gives absolutely no indication of how much uncertainty is associated with the estimate. Any sample is subject to sampling error due to variability in the population. This variation can be found by comparing results 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 that 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 to ensure that the true population means lies within the interval.

B. The statistical method for indicating reliability of a sample mean involves the use of confidence limits. Confidence limits for the mean of a sample express the confidence that the true population mean falls within a given interval. They are expressed in percentage probability terms, with a higher probability indicating a higher level of confidence that the true mean falls within the given interval. For example, the upper and lower confidence limits (40 versus 48) for the mean of a sample of age from the population of 700 farmers could be expressed as, the mean age of the population = 44 with 90-percent confidence. (Another way to say this is: If simple random samples were drawn for all possible samples of size n, 90 percent of the time the associated confidence intervals would contain the true mean. Note that the confidence intervals would differ with different samples.)

C. This statement suggests that we are 90-percent certain that the true population mean is between 40 and 48. Conversely, the interval will not contain the population mean in 1 out of 10 samples. The most common percentages of confidence used in this method are 90 percent, 95 percent, and 99 percent. Since 90-percent confidence is the same as 10-percent error and 95-percent confidence is the same as 5-percent error, etc., the most common probabilities of error are 10 percent, 5 percent, and 1 percent respectively. An interval of reliability can be calculated for the mean of a sample in the following manner:

(1) Confidence limit = mean +/- (t)*standard error of the mean.
(2) Confidence limit = the estimate +/- the bound on the error of estimation, where, the bound on the error of estimation = (t)*standard error of the mean

D. The t-value is a confidence coefficient that is based on degrees of freedom and the percent confidence chosen. In general, the degrees of freedom for a sample equal the number of observations in the sample minus 1. A matrix of "t" values is given in table 610F-2. Using the age of farmers as example, sample size is 25, so degrees of freedom is 24, if 95-percent confidence is chosen as the level of reliability.

E. The upper and lower bounds on the confidence intervals are calculated as the estimate +/- the bound on the error of estimation. Examples of calculating confidence limits for the estimates of the population mean, total, and proportion are included in the sections 610.57, 610.58, and 610.59, respectively.

F. In table 610F-2, with degrees of freedom, and probability of error (or percent of confidence) required, "t" can be obtained from moving down the .05 probability of error (95-percent 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-value. Given the mean age as 47.4 and the standard error as 2.71, the confidence limits could be derived as follows:

(1) Confidence limits = 47.4 +/- (2.07) (2.71) = 47.4 +/- 5.6
(2) Confidence limits = 41.8 and 53.0

G. Thus, one can be 95-percent sure that the average age of the 700 farmers falls between 41.8 and 53, given the mean of the 25 farmer sample.

Table F-2 t-Table

Freedom	**50 (50%)**	**.10 (90%)**	**.05(95%)**	**0.1 (99%)**
1	1.000	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	1.65	1.98	2.61
200	.675	1.65	1.97	2.60
300	.675	1.65	1.97	2.59
400	.6.75	1.65	1.97	2.59
500	.674	1.65	1.96	2.59
1000	.674	1.65	1.96	2.58
1000+	.674	1.64	1.96	2.58

H. Another example: Determine the 90-percent confidence limits for the mean of the education level of the farmers in the conservation tillage. Given the mean of the 25-farmer sample as 12.8 years of education with a standard error of .65, then the t-value from the t-table is 1.71 (under the .10 (90-percent) column of the t-table, that is interpolated between 1.72 and 1.70. Substitute these values into the equation to derive the confidence limits, as follows:

(1) Confidence limits = 12.8 +/- (1.71) (.65) = 12.8 +/- 1.1
(2) Confidence limits = 11.7 and 13.9.

I. Thus, one can be 90-percent sure that the average education level of the 700 farmers falls between the near-high-school-graduate level and the 2-year college level, given the mean of the 25-farmer sample.

610.56 Hypothesis Testing

A. Introduction

(1) A hypothesis can be defined as a tentative theory or supposition. Everyone hypothesizes from time to time when observations are made. For example, the following statements could be taken as hypotheses:
(i) The average height of American adult males is 5 feet, 9 inches.
(ii) The soybean yield in watershed X is 35 bushels per acre.
(iii) The average row crop farmer in the Midwest conservation tills half his cropland.
(2) These three hypotheses are statistical hypotheses because they are statements about a statistical population; specifically, they are statements about certain variables (characteristics such as height, yield, and percent of conservation tillage) in a statistical population.
(3) It is often desirable to test if such hypotheses are valid. To do this, an appropriate sample is taken, and the hypothesis is rejected or not rejected based on the results of statistical tests.

B. Testing the Mean

(1) The average 600-acre farmer in Cook and Haynes counties is 55 years old, has an eighth grade education, and conservation tills 200 of his acres. These are three separate hypotheses about the same 700-farmer population. How would one go on testing whether these estimates are accurate? One option 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. Instead, one could use the following process:
   (i) Interview an appropriate-sized random sample of the 700 farmers.
   (ii) Gather information about age, education, and degree of conservation tillage used.
   (iii) Calculate the sample mean for each of the three variables.
   (iv) Use hypothesis testing methods to compare the sampled means to the three hypotheses made above.

(2) Assuming steps i, ii, and iii have already been completed (using hypothetical data) and the results of these steps helped calculate the sample means, step iv involves the use and comparison of two "t" statistics: the "t-value" shown in the t-table (just as was done for confidence intervals,) and the "calculated t," which is calculated using information from the sample.

(3) As was done to calculate confidence intervals, the percent confidence (e.g., 90 percent, 95 percent, or 99 percent) required of the test must be determined. Using this value, plus the degrees of freedom, the proper t-value can be found in the t-table. In this case, the sample size is 25, so the degrees of freedom are one less than that, or 24. If a 90-percent confidence is required in the hypothesis test, the next step is to interpolate between 20 and 30 degrees of freedom in the t-table, under the .10 (90-percent) column. The proper t-value is equal to 1.71.

(4) The equation used to find the calculated t is the absolute difference between sample mean and hypothesized population mean divided by the standard error of the sample mean.

   Calculated t = |sample mean – hypothesized mean| / standard error of the sample mean

(5) Given the sample mean and the standard error of the sample mean, the calculated t-value for the farmers' age is derived by subtracting the hypothesized mean (55) from the sample mean (47), which gives -8. The absolute value |-8|=8 is then divided by the standard error of the mean, 2.71, to yield a calculated t-value of 3.86:

   The calculated t-value for the farmers' age = |47-55| / 2.71 = 3.86

(6) The comparison rule for t-value from the t-table and the calculated t is as follows:
   (i) If the calculated t exceeds t –value from table 610F-1, reject the hypothesis.
   (ii) If not, do not reject the hypothesis.

(7) The age hypothesis stated that the 700 farmers' average age is 55. The calculated t (3.86) is greater than the t-value from the t-table (1.71), so this hypothesis is rejected. (The sample tends to show the average age is less than 55.) The test is to reject with 90-percent confidence, thus a 10-percent margin of error is acceptable.

(8) 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. If the 90-percent level of confidence is satisfactory, the t-value in the t-table would be the same as in the same example (t=1.71). The calculated t in this example would be:

   The calculated t for education = |12.8 – 8| / 0.65 = 7.38

(9) In the equation, 12.8 is the given sample mean, 8 is the hypothesized mean, and 0.65 is the given standard error of the mean. The calculated t (7.38) is larger than the t-value in the t-table (1.71), 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 average.)

(10) The third test is on the hypothesis that the average farmer conservation tills 200 of his 600-acre farm. Assuming a 90-percent confidence level again, the t-value in the t-table is 1.71. The calculated t is .59 using the given data from the sample under the conservation till variable as below:

The calculated t for conservation tilled acres = $|220 – 200| / \div 34.16 = 0.59$.

(11) The calculated t (0.59) does not exceed the t-value in the t-table (1.71), so the hypothesis that the average farmer conservation tills 200 acres is not rejected. To summarize, statistical procedures were used to test three different hypotheses regarding 700 farmers in Cook and Haynes counties. The procedures involve comparison of a value (calculated t), which is derived using the combined results of the sample and the hypothesis to a t-value (in the t-table), and a confidence constant, which takes into consideration the sample size and an acceptable level of precision.

(12) The failure to reject 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; confidence in this decision, however, is only as high as the level of confidence used to obtain the t-value in the t-table.

(13) If we have two different hypotheses about the true mean we need to test them to find out which has the greater probability of being true. When we test a null hypothesis, H0, against an alternative hypothesis, H1, based on sample observations, two types of error may occur:

Decisions/Facts	H0 is true	H0 is false
Fail to reject H0	correct decision: true negative	Type II error: false negative
Reject H0	Type I error: false positive	correct decision: true positive

(14) The probability of making type I error is called α and the probability of making type II error is called β. Though we cannot eliminate the errors completely, we may control the probabilities of making these errors by collecting an appropriate number of samples and by setting appropriate critical regions (of two normal distribution curves) (Chu, 1968).

C. Test for Difference Between Two Means – Comparing Two Groups of Data

(1) One of the most-often used statistical hypothesis tests is the test for difference between two means. The differences being investigated are those between two means, and the hypothesis being tested states that the two means are equal. Uses for this test could include comparing the rate of weight gain of hogs on two different rations, comparing math scores of male verses female students, or testing to see if the 300 farmers in Cook County vary from the 400 in Haynes County in terms of age, education, and the use of conservation tillage. The actual hypothesis would state that

there are no differences in average scores between male and female students; and that there are no differences between average age, education, or use of conservation tillage between farmers in the two counties, respectively.

(2) To test each of the three hypotheses, that there is no difference between average use of conservation tillage, age, and education between farmers sampled in Cook and Hynes counties, a "calculated t" and a t-value from the t-table are compared to determine whether each of the three hypothesis is rejected or not, just as was done for the "Testing the Mean" procedure. However, the calculated t formula for the test for difference between two means is much more complicated than the one for testing the mean (see appendix of formulas). The decision rule for the test for difference between two means is: If the calculated t is larger than the t-value from the t-table, reject the hypothesis (this is the identical rule used in testing the mean).

(3) The Excel Data Analysis ToolPak has one t-Test for paired samples and two t-Test for unpaired samples. One t-Test for unpaired samples assumes that the two samples from populations with equal variances and the other do not. With the equal variance assumption, the t-Test uses the pooled variance of the two samples in the t-Test formula; without the equal variance assumption, the t-Test uses the average of the two sample variances in the t-Test formula. The data about the farmers of Cook and Hynes counties are unpaired samples.

(4) To test the hypothesis that the average use of conservation tillage is the same in Cook and Haynes counties using the Excel t-Test function, you need to first sort the data set by county, and then click the "Data Analysis" button in the "Analysis" group on the "Data" tab. The next step is to choose the "t-Test: Two-Sample Assuming Equal Variances" option from the list of data analysis tools. As showing in figure 610F-6.x, in the t-Test dialog window, you need to specify the locations of inputs and output, and set the hypothesized mean difference as 0. Assuming 95-percent confidence is required in the test, you need to set the Alpha value to 0.05, then click "OK."

(5) The test output shows that the calculated t is 0.54, with 23 degrees of freedom in this example. With 95-percent confidence, the t-value is 2.07. The calculated t is not larger than the t-value derived from the t-table, so the hypothesis is not rejected. 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 proportion of their farms as do the 400 farmers in Haynes County. A t-Test with unequal variances assumption can be performance in similar fashion with little difference in results in this case.

Figure F-6 t-Test for Two Unpaired Samples

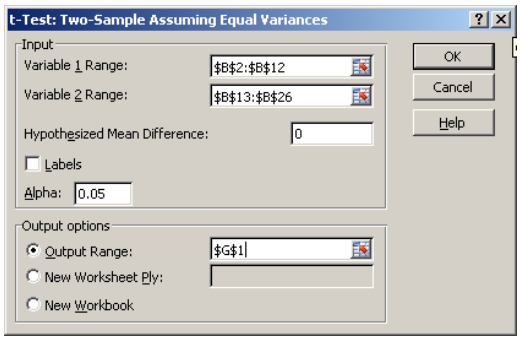
OBS	ACRES	AGE	EDUC	COUNTY
1	190	50	14	Cook
5	340	39	12	Cook
7	210	51	14	Cook
9	55	67	8	Cook
12	0	68	8	Cook
14	80	59	12	Cook
15	600	30	18	Cook
19	480	31	16	Cook
20	180	52	12	Cook
24	225	46	12	Cook
25	295	37	16	Cook
2	135	59	14	Haynes
3	275	39	16	Haynes
4	185	49	14	Haynes
6	575	32	16	Haynes
8	95	55	8	Haynes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

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	241.3636364	203.7857143
Variance	32485.45455	28210.64286
Observations	11	14
Pooled Variance	30069.25663	
Hypothesized Mean Difference	0	
df	23	
t Stat	0.537850742	
P(T<=t) one-tail	0.297923373	
t Critical one-tail	1.713871517	
P(T<=t) two-tail	0.595846747	
t Critical two-tail	2.068657599	

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	241.3636364	203.7857143
Variance	32485.45455	28210.64286
Observations	11	14
Hypothesized Mean Difference	0	
df	21	
t Stat	0.533126423	
P(T<=t) one-tail	0.299771703	
t Critical one-tail	1.720742871	
P(T<=t) two-tail	0.599543407	
t Critical two-tail	2.079613837	



(6) 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 using the same Excel tool. Assuming equal variances, the calculated t is 0.25. If a 95-percent confidence was assumed, the critical t value would be 2.07. Since the calculated t is not larger than the t-value from the t-table, the original hypothesis about the average age of the farmers is not rejected. Using the same procedures, it would also be concluded that the average education level between counties does not differ statistically for the total populations, since the calculated t (0.14) is less than the critical t value (2.07).

(7) The test for difference between two means is a useful technique in statistics. Many comparisons of large populations can be made using this test along with sampling techniques. The test is straight forward using the Excel Data Analysis ToolPak.

610.57 Estimating the Population Mean

A. Formulas

(1) Agronomists may want to test a new variety of seeds to determine the average number of days between planting and germination or between germination and maturity. A feed company may want to test the average weight gain for chicks fed a new ration during their first 3 weeks. A farmer may want to know the average amount of protein per ton of silage harvested on the farm. The estimators in this section provide a method for estimating the population mean from data collected as a simple random sample of the population. For a simple random sample of size n from a population of size N, estimates of population means, estimated variance of the means, and bound of the error of the estimation can be computed with the following equations:

(2) Estimator of the Population Mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \text{(F-5)}$$

(3) Estimated Variance of the Population Mean

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right), \qquad \text{(F-6)}$$

Where s^2 is the sample variance defined in equation F-2 as $s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ and the quantity $\left(\frac{N-n}{N}\right)$ is the finite population correction factor (FPC). The FPC decreases the size of the variance, thus making the estimate more precise. As a rule of thumb, if the FPC is greater than 0.95 (or equivalently if n is one-twentieth of N or smaller, the FPC can be ignored.

(4) Bound on the Error of Estimation

$$B = t\sqrt{\hat{V}(\bar{y})} = t\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}, \qquad \text{(F-7)}$$

Where t = the Student's t value. Table 610F-2 of t-values is includes earlier in this subpart.

(5) Sample Size Required

The sample size required to estimate the true mean with a bound on the error of estimation B is:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} \qquad \text{(F-8)}$$

Where $D = \frac{B^2}{4}$,

B is defined in equation F-7, and σ^2 is the population variance. We seldom know the population variance. If we conduct a pilot test survey, we can use s^2, the sample variance from that sample to estimate σ^2. Alternatively, we can use the following approximation to represent the standard deviation:

$$\sigma \approx \frac{range}{4} \qquad \text{(F-9)}$$

where the range is the difference between the lowest and highest values.

B. Examples

(1) We will use the same earthworm example used earlier to demonstrate how we can use Excel to calculate an estimate of the population mean and the estimated variance of the mean. In the random sample of 10 square-foot plots (6 inches deep) within a garden, the number of earthworms counted in each plot was 2, 5, 6, 6, 7, 9, 10, 11, 12, and 15. The sample size, n, is 10 and the population size, N, is 200.

(2) Use the "Data Analysis" button in the "Analysis" group on the "Data" tab. Select "Descriptive Statistics" from the choice list, following the same procedure as before, but check the box next to "Confidence Level for Mean" as well (Note the box to the right of "Confidence Level for the Mean" has 95 percent by default.), and hit "OK" (figure 610F-7). The set of descriptive statistics calculated for the 10 selected numbers include mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, count, and confidence level, as shown in figure 610F-8.

Figure F-7 Estimation of Population Mean Using EXCEL

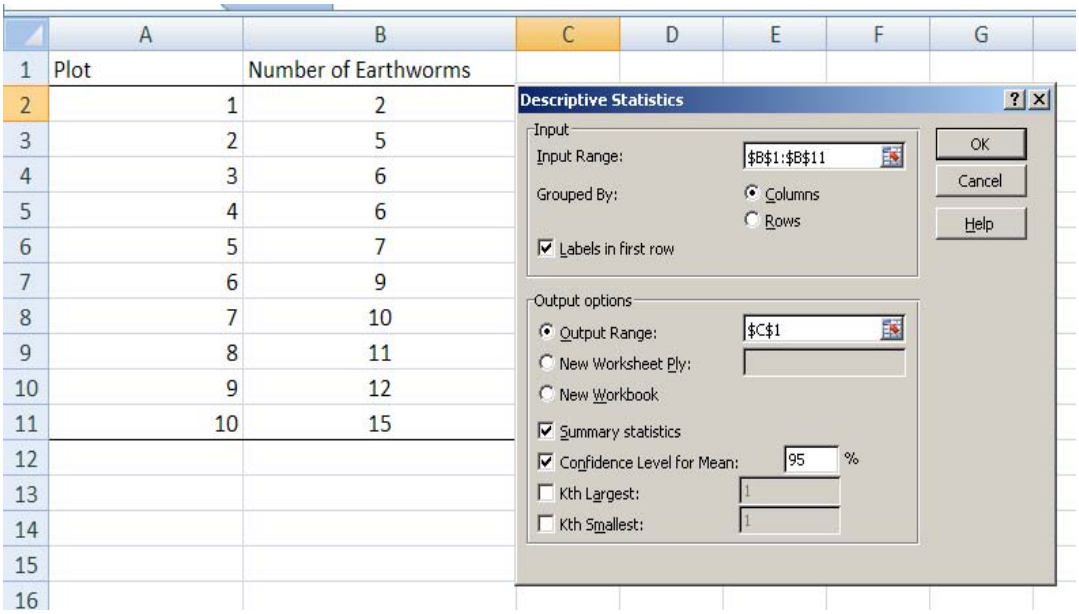


Figure F-8 Results of Estimation of Population Mean Using EXCEL

	A	B	C	D
1	Plot	Number of Earthworms	*Number of Earthworms*	
2	1	2		
3	2	5	Mean	8.3
4	3	6	Standard Error	1.21
5	4	6	Median	8
6	5	7	Mode	6
7	6	9	Standard Deviation	3.83
8	7	10	Sample Variance	14.68
9	8	11	Kurtosis	-0.27
10	9	12	Skewness	0.16
11	10	15	Range	13
12	Population Size (N):	200	Minimum	2
13	Sample Size (n):	10	Maximum	15
14	Estimated Variance of the Mean:		Sum	83
15	with fpc:	1.39	Count	10
16	w/o fpc:	1.47	Confidence Level(95.0%)	2.74
17	Estimated Std. Error of the Mean:		Confidence Level with fpc:	2.67
18	with fpc:	1.18	Upper Bound:	11.04
19	w/o fpc:	1.21	Lower Bound:	5.56

(3) Estimated Population Mean and Variance

As in figure 610F-8, the average number of earthworms per square foot in the top 6-inch layer of soil for the garden, $\overline{y}$ (equation F-5) is 8.3 worms per square foot (cell D3). A little math is required to obtain the estimated variance of the mean (equation F-6). The sample variance, s^2, is reported in cell D8 as 14.68 (rounded to two decimal places). The FPC in our example is (200-10)/200 = .95. To calculate the estimated variance of the mean that includes the FPC, type the following in cell B15: =(D8/10)*(200-10)/200. The value, rounded to two decimal places, is 1.39. Because the FPC is .95, by the rule of thumb given above we could have calculated the estimated variance of the mean without including the FPC by typing the following in cell B16: =Var(B2:B11). The resulting value, rounded to two decimal places, is 1.47. Notice that the estimated variance is larger when it is calculated without the FPC. We want our estimates to be precise (have smaller variance). If we take the square root of the estimated variance calculated without the FPC, 1.47, the result is the standard error, 1.21, reported in cell D4. The units of the standard error in this case are earthworms per square foot in the top 6-inch layer of soil for the garden, the same as that of the mean.

(4) Confidence Limits

(i) Next we will calculate the confidence limits (also called certainty limits) of our estimate of the mean number of earthworms per square foot of garden soil (6-inch depth). As shown in cell D16, the confidence limits are 2.74 for the 10 worm sample at 95-percent confidence level (you can get the same result using equation F-7). A 95-percent confidence level is the default in Excel. This is a commonly reported level of confidence that can be changed by typing in 90 or 99 in the "Confidence Level" box in the Descriptive Statistics Wizard. Cell D16 contains the 95-percent bound on the error of estimation (equation F-7) without inclusion of the FPC. The FPC may easily be included by multiplying the value

in D16, 2.74, by the square root of the FPC. Type =D16*((200-10)/200)^0.5) in D17.

(ii) We stated earlier that an FPC with a value of 95 percent or greater can be ignored. In this case, the application of FPC changes the bound of the error of estimation to 2.67 earthworms per square foot of garden soil (6-inch depth). The bound on the error of estimation, 2.74, is subtracted and added to the estimated population mean, 8.3, to get the lower and upper confidence limits of 5.56 and 11.04, respectively. This means that we are 95-percent certain that the true number of earthworms per square foot of garden soil (6 inches deep) is between 5 and 11 earthworms.

(5) Sample Size

(i) Suppose we wish to estimate the average tract size of land currently enrolled in a specific conservation program for a 10-county area. Paper records for the 800 tracts in the 10-county area currently enrolled in the program are filed by tract number. Because the records are not available electronically, we want to draw a simple random sample of the records to estimate the average tract size. We know that the sizes of the enrolled tracts range in size from 20 to 600 acres. How many of the 800 records would we have to sample (with a simple random sample design) to estimate the average size of the enrolled tracts with a bound of error of estimation B=20 acres? The range of enrolled tract sizes is 600-20 = 580 acres.

(ii) The standard deviation can be estimated as: $\sigma = \frac{\text{range}}{4} = \frac{580}{4} = 145$

(iii) The population variance, σ^2, is estimated by squaring 145 to get 21,025.

$$D = \frac{B^2}{4} = \frac{20^2}{4} = 100$$

(iv) Putting all the information into equation F-8 we get:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = \frac{800(21025)}{799(100) + 21025} = 166.7$$

(v) We should take a simple random sample of 167 records to estimate the average size of the enrolled tracts to be within 20 acres of the error of estimation.

610.58 Estimating the Population Total

A. Formulas

(1) Sometimes we may be more interested in estimating the population total than the population mean. A woodlot owner may want to estimate the total volume of wood that is ready for harvest by selecting a random sample of plots in the woodlot and taking measurements on all the trees within the selected plots. A ginseng farmer may take a random sample of 2-foot sections of rows to estimate the total weight of the upcoming harvest. Estimators in this section incorporate the population size to arrive at an estimate of the total. For a simple random sample of size n from a population of size N, estimates of population totals, estimated variance, and bound on the error of estimation may be computed with the following:

(i) Estimator of the Population Total

$$\hat{T} = N\bar{x} = \frac{N\sum_{i=1}^{n} x_i}{n} \qquad \text{(F-10)}$$

(ii) Estimated Variance of the Population Total

$$\hat{V}\left(\hat{T}\right) = \hat{V}\left(N\bar{x}\right) = N^2 \frac{s^2}{n}\left(\frac{N-n}{N}\right), \qquad \text{(F-11)}$$

Where $s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$ is the sample variance, and the quantity $\left(\frac{N-n}{N}\right)$ is the FPC. The FPC decreases the size of the variance, thus making the estimate more precise. As a rule of thumb, if the FPC is greater than .95 (or equivalently if n is one-twentieth of N or smaller), the FPC can be ignored.

(iii) Bound on the Error of Estimation

$$B = t\sqrt{\hat{V}\left(N\bar{x}\right)} = t\sqrt{N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)}, \qquad \text{(F-12)}$$

where t = the Student's t value. The t-values are available in table 610F-2. Excel can also be used to calculate the bound on the error of estimation without accessing the tables.

(2) Sample Size Required

(i) The sample size required to estimate the true population total with a bound on the error of estimation B is exactly the same as that given in equation F-8 to estimate the true population mean with a bound on the error of estimation B:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2}, \qquad \text{(F-13)}$$

where $D = \frac{B^2}{4}$, B is defined in equation F-12, and σ^2 is the population variance.

(ii) We seldom know the population variance. If we conduct a pilot test survey, we can use s^2, the sample variance from that sample, to estimate σ^2. Alternatively, we can use the following approximation to represent the standard deviation:

$$\sigma \approx \frac{\text{range}}{4}$$

(3) Examples

The examples in this section will build on those presented in the previous section. We will again use the earthworm example to demonstrate how we can use Excel to calculate an estimate of the population mean and the estimated variance of the mean. Using the earthworm data in figure 610F-9, the sample size, n, is 10 and the population size, N, is 200. We use the Descriptive Statistics Wizard in Excel to get the set of descriptive statistics for the data as described in the last example and shown in the figure below.

Figure F-9 Example For Estimating Population Mean Using EXCEL

	A	B	C	D
1	Plot	Number of Earthworms	*Number of Earthworms*	
2	1	2		
3	2	5	Mean	8.3
4	3	6	Standard Error	1.21
5	4	6	Median	8
6	5	7	Mode	6
7	6	9	Standard Deviation	3.83
8	7	10	Sample Variance	14.68
9	8	11	Kurtosis	-0.27
10	9	12	Skewness	0.16
11	10	15	Range	13
12	Population Size (N):	200	Minimum	2
13	Sample Size (n):	10	Maximum	15
14	Estimated Polulation Variance:		Sum	83
15	with fpc:	55775.56	Count	10
16	w/o fpc:	58711.11	Confidence Level(95.0%)	2.74
17	Estimated Population Std. Error:		Population Total:	1660
18	with fpc:	236.17	Population Total Bound:	
19	w/o fpc:	242.30	with fpc:	534.25
20			w/o fpc:	548.13

(4) Estimated Population Total and Variance

(i) As in figure 610F-9, the average number of earthworms per square foot in the top 6-inch layer of soil for the garden $\bar{x}$, (equation F-5) is 8.3 worms per square foot (cell D3). In this example we are interested in estimating the population total (i.e., the total number of worms in the top 6 inches of soil in the example garden). Using equation F-9, we multiply the estimated mean 8.3 worms per square foot (cell D3) by N=200 square feet in the garden to get 1,660 worms in the top 6 inches of garden soil.

(ii) A little math is again required to obtain the estimated variance of the population total (equation F-10). The sample variance s^2, is reported in cell D8 as 14.68 (rounded to two decimal places). The FPC in our example is (200-10)/200 = .95. To calculate the estimated variance of the total that includes the FPC, type the following in cell B15: = (200^2)*(D8/10)*(200-10)/200. The value, rounded to two decimal places, is 55,775.56. Because the FPC is .95, by the rule of thumb given above we could have calculated the estimated variance of the mean without including the FPC by typing the following in cell B16: (200^2)*(D8/10). The resulting value, rounded to two decimal places, is 58,711.11. Notice that the estimated variance is larger when it is calculated without the FPC. We want our estimates to be precise (have smaller variance). If we take the square root of the estimated variance calculated without the FPC, 58,711.11, the result is the standard error, 242.30. The units of the standard error in this case are earthworms in the top 6-inch layer of soil for the garden, the same as that of the total.

(5) Confidence Limits

(i) Next we will calculate the confidence limits (also called certainty limits) of our estimate of the total number of earthworms in the garden soil (6-inch depth). As shown in cell D16, the confidence limits are 2.74 for the 10-worm sample at 95-percent confidence level (you can get the same result using equation F-7). In this example we are interested in computing the error of estimation for the total (equation F-12). To get this estimate without the inclusion of the FPC, we

multiply the values in D20 by N=200 to get 548.13 worms in the top 6 inches of garden soil. The FPC may easily be included by multiplying the value in D16, 2.74, by both N and the square root of the FPC (equation F-12). Type =200*D16*((200-10)/200)^0.5 in D19.

(ii) The result is 534.25 earthworms per square foot of garden soil (6-inch depth). The bound on the error of estimation, 534, is subtracted and added to the estimated population total, 1660, to get the lower and upper confidence limits of 1126 and 2194, respectively. We are 95-percent certain that the total number of earthworms in the garden soil (6 inches deep) is between 1,126 and 2,194 earthworms.

(6) Sample Size

(i) The equation used to estimate the sample size needed to estimate the total with a bound on the error of estimation (equation F-13) is the same as that used to estimate the sample size needed to estimate the means with a bound on the error of estimation (equation F-8).

(ii) In the example in section 610.56, we wanted to estimate the average tract size of land currently enrolled in a specific conservation program for a 10-county area. Paper records for the 800 tracts in the 10-county area currently enrolled in the program are filed by tract number. Because the records are not available electronically, we wanted to draw a simple random sample of the records to estimate the average tract size. We knew that the sizes of the enrolled tracts range in size from 20 to 600 acres and wanted to determine the size of the sample required to estimate the average size of the enrolled tracts with a bound on error of estimation B=20 acres.

(iii) In this section instead of estimating the sample size needed to estimate the average tract size that has a bound on error of estimation B=20 acres, suppose we would like to determine the sample size required to estimate the total number of acres enrolled in the conservation program to have a bound on the error of estimation B=100. All the calculations would be the same as that of section 610.56, except that we would substitute B=100 for B=20 in the example.

$$\sigma = \frac{\text{range}}{4} = \frac{580}{4} = 145$$

- The population variance, σ^2, is estimated by squaring 145 to get 21,025.

$$D = \frac{B^2}{4} = \frac{100^2}{4} = 2500$$

- Putting all the information into equation F-8 we get:

$$n = \frac{N\sigma^2}{(N-1)D + \sigma^2} = \frac{800(21025)}{799(2500) + 21025} = 8.3$$

- We should take a simple random sample of nine records to estimate the total acreage of the enrolled tracts to be within a bound of error of estimation of 100 acres. If we wanted to estimate the total acreage of enrolled tracts to within 20 acres of error, we would have used B=20 and would have needed a sample size of 167 as in section 610.56.

610.59 Estimating the Population Proportion

A. Formulas

(1) Sections 610.56 and 610.57 required ratio scale or interval scale data for completing the desired calculations. This section will present methods of estimating the

population proportion and the estimated variance of the population proportion for nominal scale data (e.g., proportion of the population that is 18 years of age or older, the proportion of voters who support a particular candidate, or the proportion of tracts registered for a particular conservation program that are in compliance.) We will consider only the case of two classes. The data are often coded as 1 if the element sampled has the characteristic of interest or 0 if it does not. Using the descriptive statistics in the data analysis tools of Excel to estimate variance and confidence intervals for this data will produce erroneous results.

(2) For a simple random sample of size n from a population of size N, estimates of population proportion, estimated variance of the proportion, and bound of the error of the estimation can be computed with the following equations based on Cochran (1977) and Schaeffer et al. (1996):

(3) Estimator of the Population Proportion

$$\hat{p} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{(F-14)}$$

(4) Estimated Variance of the Population Proportion

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right) \qquad \text{(F-15)}$$

where $\hat{q} = 1 - \hat{p}$ and the quantity $\left(\frac{N-n}{N}\right)$ is the FPC. The FPC decreases the size of the variance, thus making the estimate more precise. As a rule of thumb, if the FPC is greater than 0.95 (or equivalently if n is one-twentieth of N or smaller, the FPC can be ignored.

(5) Bound on the Error of Estimation

$$B = t\sqrt{\hat{V}(\hat{p})} = t\sqrt{\frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)}, \qquad \text{(F-16)}$$

where t = the Student's t value with n-1 degrees of freedom. See table 610F-2 for a table of t-values. Excel cannot be used to directly calculate the bound on the error of estimation for nominal data. We can approximate a 95-percent bound of error by substituting t=2 in equation F-16.

(6) Sample Size Required

The sample size required to estimate the true mean with a bound on the error of estimation B is:

- $$n = \frac{Npq}{(N-1)D + pq} \qquad \text{(F-17)}$$

where $D = \frac{B^2}{4}$, B is defined in Equation F-7, and q = 1 - p. We seldom know p, but can sometimes estimate it from past surveys. If we have no prior information, we can use p = 0.5 as a conservative estimate.

B. Examples

(1) Estimated Population Proportion and Variance

(i) Suppose we want to determine the proportion of tracts enrolled in a conservation

program within a region that are in compliance of the agreement. A simple random sample of size n=80 is drawn from a total N=2000. The 80 tracts were checked and 72 were found to be in compliance. An estimate of the proportion of the 80 tracts that are in compliance is 72/80 = 0.9. We estimate that 90 percent of the enrolled tracts are in compliance.

(ii) The estimated variance of the population proportion estimate may be calculated using equation F-15.

- $$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right) = \frac{(.9)(.1)}{79}\frac{(2000-80)}{2000} = .00109$$
- The standard error of the estimated proportion is the square root of the estimated variance, 0.033.

(2) Confidence Limits

The bound of the error of estimation is calculated with equation F-16. Notice that this is equivalent to multiplying the standard error of the estimated proportion by 2.

- $$B = t\sqrt{\frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)} = 2\sqrt{\frac{(.9)(.1)}{79}\frac{2000-80}{2000}} = 0.066$$
- The bound, 0.066 is subtracted and added to the estimated proportion, .90, to get the lower and upper confidence limits (0.834 and 0.966), respectively. We are 95-percent certain that the true proportion of tracts that are in compliance is between 83.4 percent and 96.6 percent.

(3) Sample Size

(i) Suppose we want to determine the proportion of a potato field that is infested with potato beetles. The field is 100 feet in length and is planted with 25 rows of potatoes. We will examine the underside of leaves on plants in 2-foot sections of the rows to determine the presence or absence of one or more clusters of eggs. We decide to take a simple random sample of the 1,225 2-foot sections of rows in the field. How many 2-foot sections do we need to sample to estimate within 3 percent of the error of estimation the proportion of the field that is infested with potato beetles?

- In this example, N=1,225 and B=0.03.
- $$D = \frac{B^2}{4}\frac{0.03^2}{4} = 0.000225$$
- Since we do not know the proportion, p, we will use p=0.5 in equation F-17.
- $$n = \frac{Npq}{(N-1)D+pq} = \frac{(1225)(.5)(.5)}{(1224)(0.000225)+(.5)(.5)} = 582.9$$

(ii) A simple random sample of 583 2-foot sections of field would be needed to estimate the infested proportion of the field to within 3 percent of the error of estimation. If instead we wanted to estimate the infested proportion to within 5 percent, a simple random sample of 302 2-foot sections would be required.

610.60 Linear Regression Analysis and Prediction

A. Introduction

(1) In the previous discussion, except for the correlation coefficients, all of the problems considered involved only one 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 observations in different samples were compared, but generally this comparison was based on only one measurement or variable per

comparison. Statistical inference analysis will be based on two or more variables of a sample. For example, more adequate judgments about farmers' use of conservation tillage can be made if characteristics that may affect this use, such as their age or education level, can be studied simultaneously.

(2) Linear regression analysis is concerned with the relationships between two 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 between the farmers' ages and education levels to their use of conservation tillage.

(3) 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:

(i) $Y = Bo + B1X1 + B2\ X2 + \ldots + Bn\ Xn$ (F-18)

(ii) Where:

- Y = Variable to be explained, called the dependent variable, e.g., use of conservation tillage (data from a sample)
- Bo =Intercept (to be solved)
- Xi=Variable or variables used to explain Y, called independent variables, e.g., age, education (data from a sample)
- Bi=Unknown parameters (to be solved)
- i = 1, 2, -----, N, Number of independent variables

(4) The term, linear regression model or function, stems directly from the use of the linear model, where the dependent variable has a linear relationship with any of the independent variables. If the regression function has only one independent variable and if it is plotted on a graph, the relationship will be expressed in a straight line. If there is more than one independent variable, the relationship between the dependent variable and any of two independent variables will be expressed in a flat plane.

B. Simple Linear Regression

(1) Assuming Y, a dependent variable and an independent variable X, have a linear relationship, their simple linear regression model would be represented as:

(i) $Y = Bo + B1\ X1$ (F-19)

(ii) Where:

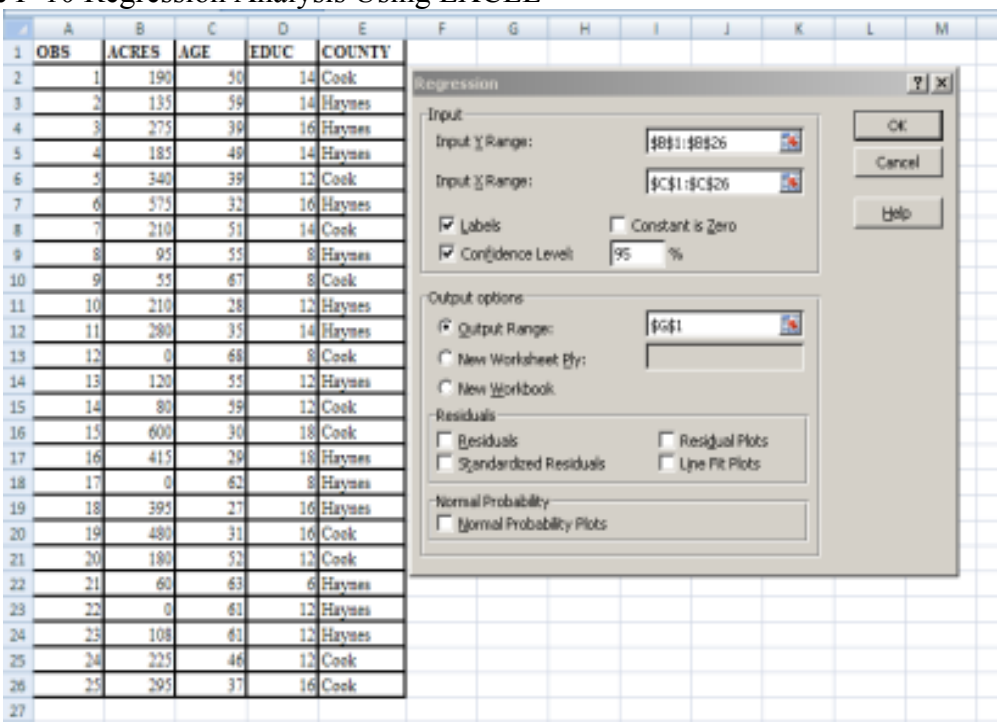
- Y = Dependent variable (from the sample)
- Bo =Intercept (to be solved)
- B1=Unknown parameters (to be solved)
- X1=Independent variable (from the sample)

(2) Assume the interest is 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? Finally, can the use of conservation tillage by similar farmers be predicted based on this information?

(3) Use sample data on the age of the farmers and their use of conservation tillage sampled in Cook and Haynes counties. The regression analysis can be conducted in Excel by using the Regression Wizard in Excel (in the "Data Analysis" button in the "Analysis" group on the "Data" tab), as shown in figure 610F-10. The regression analysis in Excel has the following procedure. First, select the "Regression Wizard"; second, type in B1:B26 as "Input Range" for Y variable and C1:C26 as "Input

Range" for X variable, check the boxes next to "Labels," "Confidence Level," and "Output Range," type in G1 as the starting point for outputs. Finally, click "OK."

Figure F-10 Regression Analysis Using EXCEL



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

Figure F-11 Regression Analysis Result Using EXCEL

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

SUMMARY OUTPUT

Regression Statistics

Multiple	0.881415
R Square	0.776893
Adjusted	0.767192
Standard	82.41994
Observati	25

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressic	1	544051.4	544051.4	80.08945	5.95E-09
Residual	23	156240.1	6793.047		
Total	24	700291.4			

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	746.9159	61.10762	12.22296	1.53E-11	620.5052	873.3267	620.5052	873.3267
AGE	-11.1096	1.241399	-8.94927	5.95E-09	-13.6776	-8.54159	-13.6776	-8.54159

(4) As shown in the regression output in figure 610F-11, the simple regression function between farmers' use of conservation tillage and their age could be expressed as follows:

Use of conservation tillage = 746.92 – 11.11 (farmers' 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 for difference between two means and testing the mean. The calculated t and degrees of freedom could be derived from the available information for testing the hypothesis that B0 = 0 and B1 = 0. If the hypothesis that B0 = 0 is not rejected, then the model becomes:
- Use of conservation tillage = – 11.11 (farmers' age)
- If the hypothesis that B1 = 0 is not rejected, the equation falls apart and it must be assumed that the age of a farmer has no effect on his use of conservation tillage.

(5) For this example, assuming 95-percent confidence with 23 degrees of freedom (minus 2 degrees of freedom, due to two parameters in the regression model) and using table 610F-2, the interpolated t-value equals 2.08. The absolute values for the calculated t's of both B0 and B1 (12.22 and -8.95, as shown in the regression output figure above) are greater than 2.81, so both hypotheses that the parameters are equal to zero are rejected. The equation is tested and remains as:

Use of conservation tillage = 746.92 – 11.11 (farmers' age)

(6) Thus, the relationship described previously between age and use of conservation tillage also stands. This equation means that 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)).

(7) For a second example, assume there is interest in whether or not the education level of farmers affects their use of conservation tillage. The regression analysis can be conducted in Excel using the Regression Wizard, with the change of the "Input Range" for x to D1:D26, and the result is presented in the figure below.

Figure F-12 Another Example of Regression Analysis Using EXCEL

OBS	ACRES	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

SUMMARY OUTPUT

Regression Statistics	
Multiple	0.80688
R Square	0.651056
Adjusted	0.635884
Standard	103.075
Observati	25

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressic	1	455928.8	455928.8	42.91312	1.1E-06
Residual	23	244362.6	10624.46		
Total	24	700291.4			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-319.86	84.99785	-3.76315	0.001011	-495.691	-144.029	-495.691	-144.029
EDUC	42.20156	6.44219	6.55081	1.1E-06	28.87488	55.52825	28.87488	55.52825

(8) The regression analysis results in the following equation:

Use of conservation tillage = -319.86 + 42.20 farmers' education

- The absolute values of calculated t's for B0 and B1 (3.76 and -6.55) are both larger than the t-value from the t-table (2.08). Thus, the hypotheses that B0 = 0 and B1 = 0 are rejected, thus the original equation remains. The above equation implies that 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-squared shown in the regression output is the estimated coefficient of determination. This figure indicates the fraction of total variation in the dependent variable that can be explained by changes in the independent variable. In the two examples, both age and education were separately found to be significant (using the test of significance) in explaining the variation of acres conservation tilled. Because the regression equation with age as the independent variable has an R-squared of 0.78, vs. 0.65 for the regression equation using education as independent variable, it did a better job of explaining the dependent variable (use of conservation tillage). Thus, the regression equation with age as the independent variable is more useful in prediction of the use of conservation tillage. (Studying the R-squared formula in the appendix of formulas will aid understanding the implications of the actual measures.)

C. Multiple Linear Regression

(1) The regression model for multiple linear regression is represented as:
(i) $Y = Bo + B1X1 + B2\ X2 + \ldots + Bn\ Xn$ (F-20)
(ii) Where:
- Y = Dependent variable (from the sample)
- Bo =Intercept (to be solved)
- B1, B2, B3, – Bn = Unknown parameters (to be solved)
- X1, X2, X3, – Xn =Independent variables (from the sample)

(2) 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: Use of conservation tillage = Bo + B1 (farmers' age) + B2 (farmers' education)

(3) The multiple regression analysis can also be conducted using the Excel Regression Wizard. The only difference from the regression analyses above is to type in C1:D26 as "Input Range" for x. The result of the regression analysis is shown in the figure below.

Figure F-13 Multiple Regression Analysis Using EXCEL

OBS	ACRES	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

SUMMARY OUTPUT

Regression Statistics	
Multiple	0.901942
R Square	0.813499
Adjusted	0.796544
Standard	77.04932
Observati	25

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regressic	2	569686.3	284843.2	47.98088	9.5E-09
Residual	22	130605.1	5936.597		
Total	24	700291.4			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	400.2317	176.344	2.269607	0.033381	34.51655	765.9468	34.51655	765.9468
AGE	-8.10914	1.85248	-4.37745	0.00024	-11.9509	-4.26733	-11.9509	-4.26733
EDUC	15.97356	7.686956	2.078009	0.049583	0.031791	31.91533	0.031791	31.91533

(4) From the regression analysis output, the relationship between conservation tillage and independent variables Age and Education can be expressed as follows: Use of conservation tillage = 400.23 – 8.11 (farmers' age) + 15.97 (farmers' education)

(5) Before this model is used for prediction, it is imperative to test the significance of each of the parameters (B's). From the regression output, the absolute values of the t values for all B's are 2.27, 4.38, and 2.08, respectively, all of which either equal or exceed the critical value of t at a 95-percent confidence level (2.08). Thus, the tests reject the hypotheses that the parameters are equal to zero and accept the model as originally specified.

(6) This model can be used to predict the use of conservation tillage by a farmer with 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 one independent variable is used.

(7) As shown below, it is obvious that the choice of independent variables is important to predicting the number of acres that are conservation tilled. For example, the predictions of conservation tillage for a 48-year-old farmer with a high school education can be different based which of the above regression equations are used:

Model	Prediction
Conservation tillage acres = Bo + B1 (AGE)	214 Acres
Conservation tillage acres = Bo + B1 (EDUC)	187 Acres
Conservation tillage acres = Bo + B1 (AGE) + B2 (EDUC)	202 Acres

(8) The test of significance will help to eliminate from consideration independent variables that do not add to the prediction; but how does one choose which independent variables to initially include in the model? The person who can safely say which independent variables should be included in the model is someone who is knowledgeable about the application of conservation tillage, or, in other words, someone who is familiar with the dependent variables. In this example, an expert in conservation tillage would be very helpful in selecting the major variables that affect a farmer's decision to conservation till. In this example, only two variables—age and education—are considered. It is possible that a farmer's financial situation or dominant soil type could be as important. Then, the financial and soil data of the sample would need to be gathered for the regression analysis and the formulation of a predictive model.

(9) To summarize, the basic steps to follow in studying the relationships between two or more variables in linear regression analysis are as follows:

(i) Consult an expert in the area being analyzed so that the major variables involved can be included for data collection in the sample, and thus used for the model.

(ii) Use correct sampling techniques to collect the relevant data on each variable.

(iii) Construct a linear regression model in the general form:

- $Y = Bo + B1X1 + B2\ X2 + \ldots + Bn\ Xn$

(iv) If only one independent variable is used, the model is a simple linear regression. If more than one independent variable is included, the model is called a multiple regression model.

(v) Use a regression analysis software, such as the Regression Wizard in Excel, to estimate the parameters (B's) their associated t stats, and R-Square of the model.

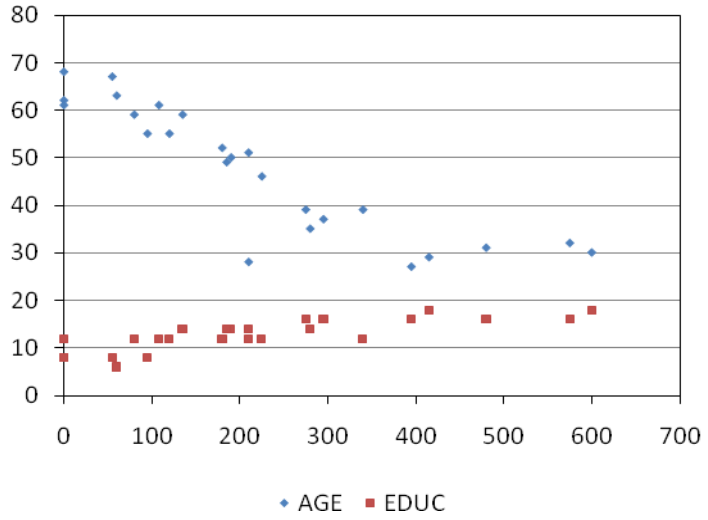
(vi) Use the test of significance (with the chosen level of confidence) to test the reliability of each parameter and, thus, its desirability as an independent variable in the equation. If the hypothesis that Bi = 0 is not rejected, then the independent "i" variable should be dropped from the equation and the remaining model should be retested.

(vii) 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 United States and Mexico are so different. When using a regression model for prediction, keep in mind the similarity of characteristics of the population from which the sample, and thus the model, were derived.

(10) Graphs are useful in the selection of independent variables that affect and help explain the variation of the dependent variable. Creating a graph similar to that shown below, for example, displaying the acres conservation tilled against the farmers' ages and education levels, could reveal the relationships between use of conservation acreage and farmers' ages and farmers' education levels.

Figure F-14 Relationship Between the Use of Conservation Acreage and Farmers' Ages and Farmers' Education Levels



610.61 Basic Statistical Formulas

Mean: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

where:

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

n = number of units in the sample

Variance: $s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$

where:

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

n = number of units in the sample

Standard Deviation: $s = \sqrt{s^2}$

where:

s^2 = Variance

Coefficient of Variation: $CV = \frac{s}{\overline{x}} \cdot 100$

where:

s = Standard Deviation

$\overline{x}$ = Mean

Standard Error of the Mean: $s_{\overline{x}} = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$

where:

s^2 = Variance of the sample

n = Sample size

N = Population size

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

Correlation Coefficient: $$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] \left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right]}$$

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

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

where:

t = t-value from Student's t distribution, Table 2

V = variance of the sample

E=acceptable error from the mean value

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

where:

R = expected range of data

CONFIDENCE INTERVALS

Confidence Limits: $$\overline{x} \pm (t)(S_{\overline{x}})$$

where:

$\overline{x}$ = mean

t = t-value, Table 2

$S_{\overline{x}}$ = standard error of the mean

HYPOTHESIS TESTING

Calculated t: $$t = \frac{\overline{x} - \mu}{S_{\overline{x}}}$$

(Test of the Mean)

where:

$\overline{x}$ = sample mean

μ = hypothesized mean

$S_{\overline{x}}$ = standard error of the mean

Calculated *t*:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{PV\left(n_1 + n_2\right)}{n_1 n_2}}}$$

(Test of the Difference)

where:

$\bar{X}_1 =$ sample mean of sample 1

$\bar{X}_2 =$ sample mean of sample 2

$n_1 =$ size of sample 1

$n_2 =$ size of sample 2

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

where:

$SS_1 =$ sum of squares from sample 1

$$SS_1 = \sum_{i=1}^{n_1} X_{1i}^2 - \frac{\left(\sum_{i=1}^{n_1} X_{1i}\right)^2}{n_1}$$

$SS_2 =$ sum of squares from sample 2

$$SS_2 = \sum_{i=1}^{n_2} X_{2i}^2 - \frac{\left(\sum_{i=1}^{n_2} X_{2i}\right)^2}{n_2}$$

SIMPLE REGRESSION ANALYSIS

Slope Parameter $\left(\beta_1\right)$: $\hat{\beta}_1 = \sum_{i=1}^{n} x_{1i} y_i \,/ \sum_{i=1}^{n} x_{1i}^2$

where:

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

$y_i = Y_i - \bar{Y}$

Intercept $\left(\beta_0\right)$: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$

where:

$\hat{\beta}_1 =$ slope parameter estimate

Coefficient of Determination: $$r = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

where:

$x_i = X_i - \overline{X}$

$y_i = Y_i - \overline{Y}$

Test $\beta_k = \beta_0$: $$t = \frac{\hat{\beta}_k - \beta_0}{SE\left(\hat{\beta}_k\right)}$$

where:

$\hat{\beta}_k =$ regression parameter estimate

$\beta_0 =$ hypothesized value

$SE\left(\hat{\beta}_k\right) =$ estimated standard error of $\hat{\beta}_k$

Chi-Square Test: $$\chi^2 = \sum_{i=1}^{n} \frac{\left(o_i - e_i\right)^2}{e_i^2}$$

where:

$o_i =$ observed value in cell i

$e_i =$ expected value in cell i

Testing $\sigma^2 = \sigma_0^2$: $$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where:

n = sample size

$s^2 =$ sample estimate of σ^2

Testing $\sigma_1^2 = \sigma_2^2$: $$F = \frac{s_1^2}{s_2^2}$$

where:

$s_1^2 =$ sample estimate of σ_1^2

$s_2^2 =$ sample estimate of σ_2^2

610.62 Statistical Glossary

A. Calculated t.—A confidence coefficient calculated from sample measures including the sample mean and standard error of the sample mean. The table t, or Student's distribution, permits the evaluation of deviations expressed in terms of standard errors for samples of various sizes, or simply to test for the significance of the difference between two means. A given standard error is divided into a difference or deviation (between two means), to obtain t as a basis for a test of significance.

B. Coefficient of Variation (CV).—A relative measure of variation used to compare the variability of data sets that are in different units. CV is derived by dividing standard error with the sample mean.

C. 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).

D. 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.

E. Correlation Coefficient.—A calculated coefficient ranging from -1 to 1 that measures the directional association between two sets of data.

F. Data.—Compiled information that can be used for analysis or computation.

G. 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.

H. F-Test.—The technique of the analysis of variance requires the comparison of two variances and a test for the significance of the difference between the calculated variances.

I Hypothesis.—An assumption subject to verification or proof.

J. Hypothesis Testing.—Using statistical procedures to reject or not reject an initial hypothesis.

K. 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.

L. Intercept.—The constant (or B0 in the equation of the text) parameter in a linear regression model that would appear as the intercept of a line on a two-dimensional graph.

M. Linear Regression Analysis.—The estimation of parameters in a relationship between two or more variables in a linear fashion.

N. Mean.—An arithmetic average, usually used to describe the average value of a particular sample.

O. Multiple Linear Regression.—Estimation of parameters in a relationship between one dependent and two or more independent.

P. Population.—The whole; the entire set of items or individuals from which a sample is drawn.

Q. Predictive Model.—The equation with estimated parameters that results from the regression analysis procedure that can be used to estimate or predict a new value for a dependent variable using values of known independent variables.

R. Presampling.—A mini sample used to roughly estimate the variance of a variable in a given population so that sample size can be calculated and applied to the principal sample when it takes place.

S. Random Sample.—A sample chosen so that each value of a variable in the population has an equal and independent chance of being collected.

T. Regression Parameters.—Known as beta coefficients (B's) in linear regression and solved by using the regression analysis procedure.

U. Sample.—A portion of the whole regarded as representative of the whole; a collection of data used to represent the population.

V. Simple Linear Regression.—Estimation of a relationship between two variables (one dependent and one independent).

W. Standard Deviation.—Dispersion of values about the mean, indicating the variation in a data set.

X. 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.

Y. 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.

Z. T (t-Value) 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 and analysis including sample size, confidence limits, hypothesis testing, and regression analysis.

AA. Table of Random Numbers.—Table with a large quantity of single digits of number arranged in a random fashion; used to facilitate the selection of a random sample.

BB. Test for Difference Between Two Means.—A test that checks for differences in two populations by statistically comparing sample means.

CC. Test of the Mean.—A test that uses the mean from a sample and probability theory to not reject or reject a hypothesis about the sample.

DD. 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.

EE. Variable.—A characteristic of a population that can be chosen for study.

FF. Variance.—The square of the standard deviation used to measure data variability.