The Evaluator's Statistical Handbook

ANALYTICAL TOOLS USING SAS

Office of Evaluation and Inspections

ACKNOWLEDGMENTS

Development of this guide has been a collaborative effort involving many different people within OEI. Each chapter was composed by members of TSAR (Technical Statistical Analytical representative). This group is composed of representatives from each of the regional offices and from central office. TSAR was formed to encourage sharing and collaboration between staff on technological and statistical issues. This handbook represents one of the many ways this group is meeting this challenge. A special thanks goes to the all the staff that helped edit this handbook, especially the staff from Region VI for their dedication to seeing this project through its completion.

CONTENTS

Introduction

Part I. Statistical Procedures

Part II. Data Manipulation and Sampling

INTRODUCTION

ANALYTICAL TOOLS USING SAS

Statistics plays an essential role in program evaluation. The application of statistical techniques and methods often increases the effectiveness of inspections and the impact of results. In a time of budget constraints, effective sampling and appropriate application of statistical techniques can save considerable inspection costs, while preserving the validity of study findings.

This handbook was prepared to provide guidance on how to use common statistical techniques. Although there are many statistical applications on the market, this volume focuses on only one of these applications - Statistical Analysis Software (SAS). SAS is an exceptionally powerful statistical and database manipulation program which is used by many Office of Evaluation and Inspections (OEI) staff, including the Technical Support Staff (TSS). SAS is available for use on the PC (windows version) or on the Health Care Financing Administration's (HCFA) mainframe.

The manual is divided into two Parts and nine chapters. Part I deals with the application of statistical procedures and testing. Part II deals primarily with identifying ways of gathering data about Medicare payments, methods for summarizing this data, and selecting samples. Each of the chapters include an introduction of why and when we use a particular type of analysis. Also included are examples from actual inspections which utilized the described procedure, as well as, the applicable SAS code used to get the desired output. The chapters end with an explanation of the output produced by the code and how it is used in a report.

Part I - Statistical Procedures

Chapters 1-6 deal with SAS procedures for data analysis, estimation, and hypotheses testing. Chapter 1 deals with looking at the data before the inspection begins to see how it is distributed. Knowing this information can help a great deal when designing the sample and defining the population. Chapter 2 describes the process for editing, coding, and tabulating survey responses in preparation for analysis. Chapter 3 explains the formula and SAS code for computing confidence intervals for means and totals. Chapter 4 explains the Chi-Square test for determining the relationship between two categorical variables. Chapter 5 describes how to test for differences between two continuous variables using the t-test. Finally, Chapter 6 concludes with a description of the statistical analysis process called regression, where a formula is developed to predict a dependent variable using one or more independent variables.

Part II - Data Manipulation and Sampling

Chapters 7, 8, and 9 deal with computations and data gathering at the beginning of the inspection process. First, Chapter 7 discusses a method for selecting a random sample from a data file. Next, Chapter 8 reviews the procedure for obtaining a complete billing history for a Medicare beneficiary. Although we may be studying only one aspect of Medicare billing (e.g., emergency transport), it is often valuable to review these services in light of all the other services received by the beneficiary. Finally, Chapter 9 concludes Part II with an explanation of how to get information about a particular Medicare procedure code that may be of interest in an inspection.

Appendix Material and Glossary

Appendix A provides an overview of basic statistical concepts, while Appendix B illustrates how to calculate a confidence interval using SAS Windows ASSIST. Applicable statistical terms are provided in the glossary.

PART I

STATISTICAL PROCEDURES

CHAPTER 1

EXPLORING DATA USING PROC UNIVARIATE

Analysis Question: How do I determine what my data looks like?

ABSTRACT

Before starting an inspection, it is often necessary to anticipate the distribution of the data you expect to get in your sample. By exploring the distribution of a prior sample, or such things as the one percent National Claims History file, you can determine whether you expect your data to be normally distributed, what the mean, median, and standard deviation of the data will be, or if outliers are expected. Answering these questions at the design phase of the study will help assure the highest quality sample. This chapter highlights how key descriptive statistics are obtained using PROC UNIVARIATE in SAS and what each statistic means.

Introduction

At the pre-inspection phase of our studies, knowing something about our data can be very useful. It can make a big difference in the quality and efficiency of our samples. One of the more common uses of this information is when we are trying to determine dollar amounts of unnecessary services for a particular HCPCS. Some common questions relative to the distribution of the data are:

1) Are the data normally distributed?
2) What are the mean, median and standard deviation of the data?
3) Are there any outliers in the data?

All of these questions and many more can be answered through the use of a SAS procedure called PROC UNIVARIATE. PROC UNIVARIATE is used when we are analyzing continuous variables such as allowed dollar amount or beneficiary age. The SAS code used for this procedure is very simple.

An example of the application of a PROC UNIVARIATE is given below. Although this example was taken from sample data from a previous inspection on Mental Health Services (OEI-02-92-00860), we usually run this procedure on the one percent National Claims History file prior to selecting the sample.

The variable used in the example to follow is called TOTALLW. TOTALLW is the amount allowed by Medicare for a beneficiary for services in the mental health area. The HCPCS services concerning us included: psychiatric diagnostic interview (90801); psychological testing with written report (90830); individual psychotherapy, 20 to 30 minutes (90843); individual psychotherapy, 40 to 50 minutes (90844); and group psychotherapy (90853) provided in a nursing home. The SAS code used to produce this output was:

SAS CODE

```
PROC UNIVARIATE DATA=BRT.MHALL;
VAR TOTLALLW;
RUN;
```

END SAS CODE

The first line of code tells SAS to invoke the UNIVARIATE procedure on the dataset named brt.mhall. The second line (VAR TOTALLW) tells SAS the continuous variable on this dataset we would like to know about. If the VAR statement is not included, SAS will produce statistics on every variable in the dataset. Be sure to include this statement, if you are only interested in one or two variables in your dataset. The RUN statement is only used in PC SAS. It is not needed if you are running SAS on the mainframe. This code produced the following output:

“MOMENTS” PORTION of PROC UNIVARIATE OUTPUT

```
                        Variable=TOTALLW


          N                          540                Sum
Wgts                    540
          Mean              559.5574          Sum
                                               302161
          Std Dev           766.9173          Variance
                                              588162.1
          Skewness          3.315512          Kurtosis
                                              17.05421
          USS                4.861E8          CSS
                                              3.1702E8
          CV                          137.0578          Std Mean

                                                        33.00286
          T:Mean=0          16.95481          Prob>|T|
                                               0.0001
          Sgn Rank                  73035               Prob>|S|
              0.0001
          Num ^= 0                    540
```

The following Table provides definitions for the “Moments” section (output)

1. *N* - the number of nonmissing observations -in this example we have 540 observations (beneficiaries) in our dataset

2. *Sum Wgts* - the sum of the weights of these observations. In this example it is equal to 540 since no weights were used.

3. *Mean* - the average

4. *Std Dev* - the standard deviation, commonly seen as the greek letter sigma (σ).

5. *Variance* - the standard deviation squared (σ^2).

6. *Skewnesss* - the lack of symmetry in a distribution. It is equal to zero for a symmetrical distribution. A distribution is said to have positive skewness (or be positively skewed) when it has a long thin tail at the right. In this example a value of 3.15512 indicates that this data is positively skewed. In contrast, a value which is negative indicates that the distribution has a long thin tail at the left.

7. *Kurtosis* - the relative peakedness or flatness of a distribution. In this example the value of 17.05421 indicates that the distribution has wider tails than the normal distribution. If this value had been less than three, the distribution would be less peaked and have narrower tails than the normal distribution.

8. *USS* - the uncorrected sum of squares. This is calculated by squaring each individual value, and summing these values. In the example above the USS is equal to 4.861E8. The E8 on the end means that you need to multiply the 4.861 by 100,000,000

9. *CSS* - the corrected sum of squares. This is calculated using the following formula: USS-(Mean * Sum).

10. *CV* - the coefficient of variation. It is computed by dividing the standard deviation by the mean and is expressed as a percent. In the above example, it is 766.9173/559.5574 or 137.06 percent. This is an indication that the data is extremely variable. The smaller the value of the CV, the less variability in the data.

11. *Std. Mean* - the standard error of the mean. This is computed by dividing the standard deviation by the square root of N.

12. *T:Mean=0* - the Student's t value for testing the hypothesis that the population mean is 0.

13. *Prob > |t|* - the probability of a greater absolute value for this t value.

14. *Sgn Rank* - the signed rank statistic also known as the Wilcoxon Test. The absolute value of the differences between paired observations are ranked. This is an alternative to the paired sample t-test.

15. *Prob > |s|* - the probability of a greater absolute value for this s value.

16. *Num ^=0* - The number of observations not equal to 0. In this example all 540 values are greater than 0.

QUANTILES

```
              Variable = TOTALLW

          Max     7567    99%     3302
100%      Q3      683.    95%     2190.
          Med     5       90%     5
75%       Q1      268     10%     1429.
          Min     119.    5%      5
50%               5       1%      70.5
                  18              47.5
25%       Range                   35
          Q3-
0%        Q1
          Mode    7549

                  564

                  146
```

The *Quantiles* portion of the output gives the maximum value in the file (100% Max), the minimum value in the file (0% Min), and also the 75th, 50th, and 25th percentiles. In this example, we can see that the largest value is 7567 and the smallest is 18. The median, or 50% percentile, is 268 meaning that one-half of the values are above 268 and the other half are below 268.

The right-hand column of this output shows other interesting information such as the 99th, 95th, etc. percentiles. This column is very useful in identifying outliers on the file. For example, since only 5 percent (about 27) of the values are over 2190 (the 95th percentile), we may want to look at all of these beneficiary records, if this were our universe. Also shown in the *Quantiles* portion of the output are the range (7567-18), the interquartile range (Q3-Q1), which is equal to 683.5-119.5 in the above example, and the mode of 146 (the most frequently occurring value on the file).

Finally, the *Extremes* portion of the output shows the five highest and five lowest values in our dataset. The two columns labeled "Obs" are the observation numbers of the values to the left.

EXTREMES

```
 Lowest      Obs      Highest
   Obs

   18       (194)       3470
  (311)
   28       (426)       3586
  (147)
   29       ( 19)       3779
  (289)
   31       ( 55)       4807
  (511)
   33       (114)       7567
  (414)
```

OPTIONS IN PROC UNIVARIATE

Among the options in PROC UNIVARIATE is the option to see graphs of your data. The code for this option is:

SAS CODE

```
PROC UNIVARIATE DATA=BRT.MHALL PLOT;
VAR TOTALLW;
RUN;
```

END SAS CODE

The PLOT option in the PROC UNIVARIATE statement produces a "stem and leaf” plot, a box plot, and a normal probability plot. The purpose of using the PLOT option is to provide a clear picture of what your data looks like. Outliers in your data become easy to see, and the distribution of the data is also very informative.

Other PROC UNIVARIATE options include (but are not limited to) a NOPRINT option which is used when the only purpose of the procedure is to create new data sets; a FREQ option which requests a frequency table consisting of the variable values, frequencies, percentages, and cumulative percentages; and a NORMAL option which computes a test statistic for the hypothesis that the input data come from a normal distribution.

Two other useful statements used in connection with the PROC UNIVARIATE are VAR and BY statements. The VAR statement specifies the variables on the file for which descriptive measures are calculated. If this statement is omitted, all numeric variables in the file will be analyzed. A BY statement is used when a separate analysis by groups is desired. When using the BY statement, it is expected that the input data set is sorted in order of the BY variables.

CHAPTER 2

EDITING, CODING, AND TABULATING SURVEY RESPONSES

Analysis Question: How do I determine how people responded to my survey questions using SAS?

ABSTRACT

After collecting data during an inspection, analysis of the data requires numerous steps. These steps include: editing questionnaires for inconsistent answers, coding individual responses to variables, determining the number of respondents for each question, and producing frequency distributions and cross-tabulation tables. This chapter is based on a standardized system of preparing data for analysis and producing frequency distributions and cross tabulations, all of which are an essential part of the inspection process.

Introduction

One of the basic things we do in OEI is conduct surveys of government officials, beneficiaries, health care professionals, and many other groups to obtain data on the programs and policies we are evaluating. To do this, we use mail questionnaires, telephone interviews, as well as other means to collect data in a focused and systematic way.

After collecting data, we should ask "How did respondents answer the survey questions?" To obtain the answer, we need to 1) edit questionnaires, 2) code individual responses, 3) determine the number of respondents (n) for each question, and 4) produce frequency distributions and cross-tabs. Coding of responses is usually done by hand. The first step, editing, can be done either in SAS, using a series of conditional statements, or by performing manual corrections and edits. The last three items are easily and quickly accomplished by using both the PROC FREQ or ARRAY statements in SAS. To illustrate these different steps, we will use an example from a recent OEI report entitled "HMO Customer Satisfaction Surveys." This inspection assessed how Medicare health maintenance organizations (HMOs) conducted customer satisfaction surveys. We used a stratified random sample of 95 HMO Medicare contracts, with the HMO contracts stratified into three groups of high, medium, and low Medicare enrollment.

Each of these HMOs were sent a mail questionnaire regarding customer satisfaction survey procedures and use of survey results. Seventy-two of the 95 sample HMOs returned completed questionnaires.

EDITING

Once survey instruments are returned by respondents, regardless of the type of guide (i.e., mail or telephone), each guide needs to be checked to ensure all answers provided are logically possible. For example, a respondent might have answered "no" to a question about whether they had ever conducted any customer satisfaction surveys, but "yes" to a question about whether they had ever conducted a customer satisfaction survey of only their Medicare enrollees. The logic inconsistency between these two responses requires the analyst to use his or her best judgement to edit the survey instrument. Otherwise, the respondent's answers should be excluded from the analysis of that question. Only if it is obvious that the respondent misread an earlier question, should answers be changed. Editing or excluding inconsistent answers is an important first step in analyzing respondent data and will help to prevent problems later in the analysis.

CODING

After surveys have been edited, as needed, coding for data entry can begin. Coding is the practice of assigning a unique code to each individual response. While codes can be any symbol or group of symbols, we have found it easiest to use individual letters or numbers, especially if the codes must be entered into a computer program. When SUDAAN (Survey Data Analysis software package) is going to be used for the analysis, variables in your dataset must be coded numerically.

It is a good idea to assign codes that are consistent throughout the survey instrument; for example, the response "yes" could always be coded as 1, and "no" could be 2. This makes entering the data into the computer quicker and easier. Consistent codes should also be used for other situations, such as when a question should have been answered by a respondent but was not (i.e. "no response," coded as 9), or when a question does not apply to that respondent (i.e. "not applicable," coded as 8). Making this distinction helps you analyze each question's response rate and helps you determine whether low response rates for certain questions resulted from respondents being ineligible to answer the question or choosing not to answer the question.

After you have analyzed response patterns, “no response” and “not applicable” responses for each question will be recoded as missing values using a short SAS program. This will prevent counting these responses in the denominator of total responses for each question.

DIFFERENT n

Different questions within the same survey instrument will often have different numbers of respondents (n). This is usually because there are some questions that respondents should have answered but did not, and because some questions did not apply to all respondents.

FREQUENCY DISTRIBUTIONS

Once editing, coding, and data entry are completed, it is important to "clean" your dataset by looking for and correcting data entry errors. Then it is possible to produce frequency distributions (a form of univariate analysis) by examining the distribution of survey responses on one variable (or survey question) at a time. In other words, frequency tables show us how many respondents gave each possible answer to each survey question.

CROSS TABULATIONS

The analysis for this inspection also involved doing cross-tabulations (usually involving two variables). This is an examination of the distribution of survey responses on two or more variables (or survey questions) simultaneously. In other words, we want to know how many respondents who answered one question one way also answered another question in a certain way. In the HMO study, we wanted to know how many of the HMOs that did not conduct Medicare-only surveys also did not include Medicare specific questions on their general surveys. Therefore, we cross-tabulated Question 18 (in this example, whether or not the HMO conducted Medicare only surveys) against Question 39 (whether or not they included Medicare specific questions on their general surveys). A combination of negative responses to both questions enabled us to determine how many HMOs did not obtain specific data about the Medicare population they were serving.

WEIGHTS

Weights are used if a complex sample design is used, but not necessary for designs using simple random sampling. Weights account for differing probabilities of the sampling units selected. Since the HMO study used a stratified random sample, it was necessary to weight the data when running frequency distributions and cross-tabulations.

Each of the three HMO sample strata (high, medium and low Medicare enrollment) were given a weight. Individual survey responses from each HMO were weighted by the value given the stratum from which they were selected. For example, HMOs in the high enrollment stratum were sampled at a disproportionately higher rate than those in the medium and low strata. Thus, in order to avoid giving a respondent from a high stratum HMO an undue influence on the survey findings, his or her responses were given a lower weight relative to those of respondents from the medium and low strata HMOs.

To compute values of the weight variable that will be added to your SAS dataset, you divide the universe of HMOs in each strata by the number of HMOs sampled in that strata.

COMPUTATION OF STRATA WEIGHTS

	Universe /	Sample =	Weight
Strata 1	13	13	1.0
Strata 2	130	70	1.857
Strata 3	42	12	3.5
Total	185	95	

Full Sample Weights

These weights are computed on the basis of how the sample was selected, not on how many responded.

Partial Sample Weights

If less than 100 percent respond to your survey, the weighted frequencies obtained in SAS will be less than the population. In this example, the respondents by strata were the following:

RESPONDENTS BY STRATA

	Sample	x	Weight	=	Responding Universe
Strata 1	13	x	1.0	=	13
Strata 2	54	x	1.9	=	103
Strata 3	5	x	3.5	=	18
Total					134

ANALYSIS

The following SAS program converts a database file to SAS, recodes variables, assigns values to the weight variable, and performs a PROC FREQ for analyzing data from the HMO study.

SAS CODE

```
1           FILENAME IN 'A:HMO.DBF';
2           PROC DBF DB3=IN OUT=HMODATA;
3           DATA HMODATA1;
4           SET HMODATA;
5           ARRAY NUM _NUMERIC_;
6           DO OVER NUM;
7           IF NUM='8' THEN NUM=.;
8           END;
9           IF STRATUM = '1' THEN WT = 1.0;
10          IF STRATUM = '2' THEN WT = 1.857;
11          IF STRATUM = '3' THEN WT = 3.5;
12          PROC FREQ DATA=HMODATA1;
13          TABLES Q18 Q18*Q39;
14          WEIGHT WT;
15          RUN;
```

END SAS CODE

Program Explanation

Lines 1 - 2. CONVERTING A DBASE FILE TO A SAS FILE. In line 1, you are telling SAS to retrieve a dBase file called "hmo" from your a: drive. Line 2 converts this database file into a SAS dataset called "hmodata."

Lines 3 - 4. CREATING ANOTHER SAS DATASET. In line 3, tells SAS that the name of your dataset will be called "hmodata1." Line 4 tells SAS that you want to create your dataset from an existing SAS dataset called "hmodata."

Lines 5 - 8. USING AN ARRAY AND DO LOOP TO ELIMINATE 8s. In line 5 SAS assigns an array called NUM (you can call the array anything you want). The command _NUMERIC_ is what tells SAS to use all numeric variables. Lines 6 and 7 tell SAS to perform a do loop in which zeros should be converted to missing values using an IF THEN statement. SAS represents missing values with a period. These missing values will then not be counted in the denominator. You end the do loop with the command "end" in line 8.

Lines 9 - 11. CREATING A WEIGHT VARIABLE. Lines 9 - 11 use an IF THEN statement to create the variable WT and assign weight values to the three strata.

Lines 12 - 15. COMPUTING FREQUENCY DISTRIBUTIONS AND CROSS-TABULATIONS. Line 12 tells SAS to run the frequency procedure for the hmodata1 dataset. Line 13 tells SAS to produce a frequency distribution on question 18 and a cross-tabulation on questions 18 and 39. When you do a cross-tabulation in SAS, you list the variable representing the row data first and the column data second (e.g., Q18*Q39). The WEIGHT command in line 14 invokes the weights you've assigned in lines 9 - 11. The PROC FREQ procedure ends on line 15 with a RUN statement.

OUTPUT

The “Frequency” column displays the weighted number of HMOs answering YES or NO to Q18, whether they have ever conducted a customer satisfaction survey of Medicare enrollees only.

FREQUENCY DISTRIBUTION

```
                              Cumulative    Cumulative
          Q18    Frequency    Percent            Frequency
Percent
__________________________________________________________

          YES    17.68        46.0          17.68              46.0

          NO     20.78        54.0       38.46        100.0
```

The “Percent” column shows the weighted percentage of HMOs that responded YES or NO. The “Cumulative Frequency” column shows the cumulative weighted number of HMOs responding YES or NO.

The responses to Q18 (whether or not the HMO conducted Medicare only surveys) are represented in the rows of the crosstab table on the next page. The responses to Q39 (whether or not the HMO included Medicare specific questions on their general surveys) are found in the columns of this same table. There are four rows in each cell, representing weighted frequency count, percent, row percent and column percent.

As you can see in the first row, only 3 (rounded down from 3.17) HMOs said YES to both questions 18 and 39 while 15 (rounded up form 14.51) said YES to question 18 but NO to question 39. Keep in mind that weights were computed to three decimal places, therefore the weighted frequencies are not in the form of integers.

The second row of the cell for YES to both questions is 8.24 percent. This is equal to the frequency of 3.17 divided by the total for the table of 38.46. The third row is the row percent, 17.93 percent (the result of dividing 3.17, the cell frequency, by 17.68, the row total * 100). Finally, the fourth number in the first cell of 81.91 percent, is the column percent. This is calculated by dividing the cell frequency 3.17 by the column total 3.87 and multiplying by 100. It is obvious from this table that not everyone answered both questions 18 and 39 (by the weighted totals of 38.46, which is considerably less than the responding universe of HMOs (N=134).

The PROC FREQ also has an option to compute a Chi-Square statistic. Refer to Chapter 4 for additional information about this statistic.

CROSS-TABULATION

```
Frequency
Percent                       Q39
Row Pct
Col Pct
                    YES       NO   TOTAL
     -----------------------------------
     YES                3.17      14.51 17.68
                        8.24      37.73 45.97
                       17.93      82.07
                       81.91      41.95
Q18
     -----------------------------------
     NO                 0.70      20.08 20.78
                        1.82      52.21 54.03
                        3.37      96.63
                       18.09      58.05
     -----------------------------------
     TOTAL                   3.87     34.59
                            38.46
                       10.06      89.94 100.00

     Frequency missing = 1.4
```

CHAPTER 3

CALCULATING CONFIDENCE INTERVALS

Analysis Question: Having analyzed data from a sample, how do I determine within what range the true population value lies?

ABSTRACT

Once we have analyzed data from a sample, we often need to determine how precise our estimates are. This information can be critical in understanding the significance of point estimates and making policy decisions. The confidence interval is determined by such factors as data variability, sample size, and how confident we want to be that the true population parameter lies within the interval. This chapter discusses the underlying theory of confidence intervals and shows how SAS can be used to calculate the confidence interval for the mean and explains how these results can be used to calculate the confidence interval for a total.

Introduction

In our reports, we typically use two types of estimates for population parameters. The first, called a point estimate, estimates a population parameter using a single sample statistic. The other estimation, called interval estimation, involves the calculation of confidence intervals. In each type of estimation, we are trying to answer the question, “What is the value of a population parameter?” When point estimation is used, the form of the answer is, “I don’t know for sure, but my best guess is (the value calculated from the sample).” When interval estimation is used, the form of the answer is something like, “I don’t know the exact answer, but I’m 95 percent confident that the true value falls within the following range” or “I don’t know the exact answer, but there is a probability of 95 percent that the true value falls within the following range.”

Confidence intervals are important because they give the reader an idea of the precision of the estimates. For example, there is a big difference between stating the estimated amount of allowed payments for unnecessary services at the 95 percent confidence level is $10 million +/- $1 million and the estimated amount of $10 million +/- $5 million. The first confidence interval is much more precise since the range is $2 million versus a range of $10 million with the second confidence interval.

The purpose of this section is to describe the method for calculating confidence intervals. For this chapter, we will limit the discussion to confidence intervals calculated for means and totals. However, the theory described can be utilized for other types of estimates (e.g., proportions). We will show how SAS can be used to make these calculations. (If SAS is unavailable, Audit's RATSTAT program can generate confidence intervals - see program documentation). However, before we discuss the application of these programs, we will first describe the underlying concepts utilized in calculating confidence intervals.

THEORY UNDERLYING CONFIDENCE INTERVAL CALCULATIONS

A confidence interval is based upon the concepts of the Central Limit Theorem and the Empirical Rule. The Central Limit Theorem says that if you have a simple random sample of *n* observations from a population with mean μ and standard deviation **STD** and if ***n*** is large, then the sample average$\bar{x}$ is approximately normally distributed with mean μ and standard deviation (STD/$\sqrt{n}$).

An important implication of the Central Limit Theorem is that for random variables from a population with a finite variance, the sampling distribution of the standardized sample mean approaches the normal distribution as the sample size *n* becomes infinite. Even if the sample values are not normally distributed, the sample average is approximately normally distributed. Since the sample average is approximately normally distributed, you can use the Empirical Rule to summarize the distribution of sample averages. The Empirical Rule says that 68 percent of the values in a normal distribution can be found within one standard deviation of the mean, 95 percent within two standard deviations of the mean, and so forth. Figure 1 illustrates this powerful concept. Using it, we can summarize the distribution of sample averages.

To estimate the population mean, a 95 percent confidence interval is very likely to contain the true population mean μ. What this means is that if we draw all possible samples of size *n* from a population, the probability is 95 percent that the true mean lies within two standard deviations of the sample. Only five percent of the time would we have failed to capture the population mean. The only way we could be 100 percent sure we calculated the true population mean would be to take a census of population.

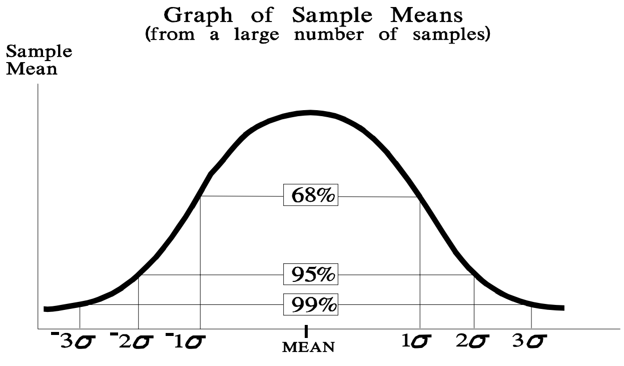


Figure 1

Standard Error of the Mean

Every statistic is derived from a sample, and although samples are defined as representative of the population from which they are drawn, they differ from each other even when drawn from the same population. Selecting a sample using a random process should generate a sample that is unbiased (no systematic differences between the sample and the parent population). We say that the deviations from the population characteristics are random, and we refer to them collectively as error variance. Now, if there is going to be error in our measurement, it is important to know the probable magnitude of that error. Otherwise, we might be overconfident in the statements that we make about the population from which a sample has been drawn. The most common way of quantifying error is to compute the standard error.

A standard error is an estimate of the standard deviation of a hypothetical distribution of values that would be obtained from a given statistic if repeated samples were drawn from a single population. For a mean, we estimate the variability of a distribution of means of successive samples of the same size from the same population. The standard deviation of that distribution is estimated by the standard error of the mean. The size of the standard error of the mean is a function of the variability of the sample and the size of the sample. If the standard deviation of the sample decreases or the sample size increases, the standard error will decrease.

Confidence Interval Formula

When the population variance is known, calculating confidence intervals is as simple as selecting the confidence level (e.g., do we want to be 95 percent or 99 percent certain that the true population mean lies within the range specified) and using the following formula:

$$\boxed{\text{Confidence Interval of the Mean} = \underset{\uparrow\atop\text{Average}}{\mu} \pm \underset{\uparrow\atop\text{normal deviate value (area under curve)}}{Z}\ \overset{\text{Population Standard Deviation}\atop\swarrow}{\sigma}/\sqrt{n} \xleftarrow{} \text{sample size}}$$

The following table lists commonly used confidence levels and corresponding z values. The greater the confidence level, the larger the value of z, making the confidence interval wider.

Level of Confidence	Z-Value
90%	*1.64*
95%	*1.96*
99%	*2.57*
99.9%	*3.30*

When the Population Variance is Unknown or the Sample is Small

In most cases, we will not know the population variance. As a result, we must use the standard deviation (s) obtained from the sample ($s/_{\sqrt{n}}$ is often referred to as the sample standard error). The formula becomes:

$$\underset{\substack{\uparrow\\ \text{Population}\\ \text{Mean}}}{\mu} = \underset{\substack{\uparrow\\ \text{Sample Average}}}{\bar{X}} \pm Z\, \overset{\substack{\sigma\text{ replaced by}\\ \text{sample standard}\\ \text{deviation}\\ \swarrow}}{s} / \underset{\leftarrow\ \text{sample size}}{\sqrt{n}}$$

Student t Distribution

The normal distribution is completely defined by μ and the population standard deviation. However, once we replace the population standard deviation with the sample standard deviation, using the normal distribution is not exactly correct. The more appropriate distribution is the student's t-distribution. Thus, the formula becomes:

$$\mu = \bar{X} \pm \overset{\substack{\text{T distribtion replaces}\\ \text{the normal distribution}\\ \text{for small sample sizes}\\ \downarrow}}{t_{df,(1-\alpha/2)}}\, s/\sqrt{n}$$

df = the degrees of freedom is one less than the sample size (df=n-1)

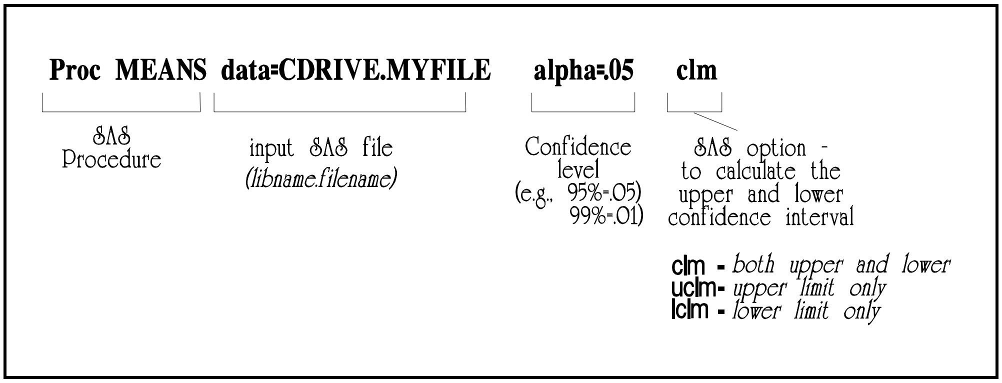
$t_{df.(1-\alpha/2)}$ is the t-value for a given degrees of freedom and alpha level. The confidence level for the interval is 1-alpha. For example, for 95% confidence level, alpha=0.05.

Remember that, when finding the t-value, you need to divide alpha by 2 before you subtract 1.

USING SAS TO CALCULATE CONFIDENCE INTERVALS

Calculating confidence intervals is extremely easy in SAS, whether on the mainframe or on the PC. The following assumes that the version of SAS software used is 6.07 or later. This version added an option to the PROC MEANS procedure which produces the upper and/or lower confidence interval.

SAMPLE SAS PROGRAM AND OUTPUT


Proc MEANS data=CDRIVE.MYFILE alpha=.05 clm
SAS Procedure
input SAS file
(libname.filename)
Confidence level
(e.g., 95%=.05)
99%=.01)
SAS option - to calculate the upper and lower confidence interval
clm - both upper and lower
uclm- upper limit only
lclm - lower limit only

SAS CODE

```
LIBNAME DDRIVE "D:\";
RUN;

PROC MEANS DATA=DDRIVE.SAMPLE94 ALPHA=.05 CLM;
VAR PAIDAMT;
RUN;
```

END SAS CODE

OUTPUT

```
Variable = PAIDAMT

Lower 95.0% CLM    Upper 95.0% CLM

  25.2367               25.7078
```

Calculating the Confidence Interval for a Total

Thus far, we have calculated the confidence interval for a mean. What if you want to estimate totals? The method for determining the confidence interval for a total includes the standard error of the mean in the calculation. These are the steps you would take:

1) Calculate the percentage added to and subtracted from the mean's point estimate.
2) Multiply the percentage obtained from step one by the point estimate of the total.
3) Determine the upper and lower bounds of the confidence interval by adding or subtracting the value calculated in step two.

An example using data from the previous output:

Given that the 95% confidence interval of the mean is 25.2367 and 25.7078 (this corresponds to the point estimate (25.4722) ± .23554):

1) Determine the difference as a percent. (0.23554/25.4722 = ±.9247%)

2) Use this percent and apply it to the point estimate of the total.

Confidence Interval of Total = Point Estimate ± (.009247) *(point estimate)

Actual Data: Say the total payment for drugs used in Texas nursing homes was $12,345,882. The confidence interval would be 12,345,882± (.009247)*(12,345,882) or 12,345,882 ± $114,162.

Lower Bound = $12,231,720
Upper Bound = $12,460,044

How to Reference a Confidence Interval in a Report

In most cases, confidence intervals for estimates are listed in the Appendix of a report. Using confidence intervals in the body of the report should be done sparingly and only to emphasize the precision or power of the estimate.

Examples of References to a Confidence Interval in the Body of a Report

- "At the 95 percent confidence level, we project Part B charges between $3.6 and $4.7 billion were made in 1992 on behalf of residents during nursing home stays."

- "For example, at the 95 percent confidence level, as few as 8.1 percent or as many as 18.1 percent of the 1,426 provider numbers identified at Empire Blue Shield had allowed charges."

Examples of References to a Confidence Interval in the Appendix

A. Wound Care Product

Allowances

		Projected Total	**Confidence Interval**
Hydrogel Dressings	Total	$32,382,970	+/- $9,270,159
	Questionable	$24,778,466	+/- $7,620,692
	Percentage	77%	+/- 4%

B. 1991 Payments for Non-Legitimate Devices (Body Jackets)

Universe Size	Sample Size	Estimated Payments	Lower 95%	Upper 95%
12,000	120	$7,021,040	$6,261,629	$7,780,450

C. National Practitioner Databank Reports

Description	**Value (%)**	**95 % Confidence Interval (+ or -)**

Proportion of reports considered useful	95.7%	2.9%

CHAPTER 4

DETERMINING RELATIONSHIPS BETWEEN CATEGORICAL VARIABLES

Analysis Question: How do I determine if there is a relationship between categorical variables?

ABSTRACT

Much of the data we collect and analyze is categorical, meaning variable responses can be classified into mutually exclusive categories such as yes/no or exhibit some kind of ranking such as 1-5 (very satisfied - very unsatisfied). Determining whether two categorical variables are independent requires the use of the Chi-Square test. A significant Chi-Square statistic shows relationships exist between variables, but will not show the direction of the relationship or whether one variable caused another to occur. This chapter illustrates an essential tool that helps analysts determine such things as whether response bias exists or which cross-tabulation tables result in statistically meaningful differences between two variables.

Introduction

In many ways, the data that we encounter in inspections can be considered categorical data. Whether our inspections look at how satisfied Medicare beneficiaries are with the service they receive or if the services provided to Medicare beneficiaries were medically necessary, we are examining data that can be classified into exhaustive and mutually exclusive categories. In addition, the data may exhibit some kind of ranking, such as degree of satisfaction.

After collecting data, we begin describing the data by counting the number of observations in each response category and computing percentages. To determine whether there are relationships between categorical variables, statisticians have developed numerous tools.

One of the most useful statistical tools for categorical data analysis is the Chi-Square Independence Test. The Chi-Square test statistic measures whether two categorical variables are independent. In other words, one variable does not tell us anything about a second variable. The size of the Chi-Square test statistic indicates whether the difference between observed and expected values is due to random error or reflects the influence of one variable on the other.

In this chapter, we present three examples from recent OEI reports of the use of the Chi-Square independence test in data analysis. The first example explains how the Chi-Square test is used to determine if there is a relationship between beneficiaries' geographic location and their feelings about services received. This example will show, step-by-step, how one calculates a Chi-Square test statistic. The second example illustrates the use of the Chi-Square independence test to determine if there is an association between respondent status and size of respondent assets. This example will show how to use SAS to perform the Chi-Square independence test for data collected for a simple random sample. The third example illustrates using a spreadsheet application to calculate Chi-Square.

Example 1 "1993 Medicare Beneficiary Satisfaction: Michigan," OEI-05-92-00390

Using Chi-Square Test to Determine Relationships between Two Variables

The Health Care Financing Administration (HCFA) began receiving complaints about the service provided by the Michigan carrier. At that time, OEI was preparing to conduct an inspection looking at Medicare beneficiaries' satisfaction with the services provided by their local carriers (OEI-04-92-00480). A decision was made to conduct a parallel survey of Medicare beneficiaries living in Michigan using the same survey instrument.

The subsequent report found that Michigan beneficiaries are more dissatisfied with the service they receive when they call their carrier. However, while the initial analysis showed that 21 percent of Michigan beneficiaries were dissatisfied, only 14 percent in the national sample were dissatisfied with the service they received when calling their carrier. We questioned whether this difference was significant enough to conclude that Medicare beneficiaries living in Michigan were more dissatisfied with the service they received than beneficiaries nationally. We used the Chi-Square independence test to determine whether the difference could be attributed to random chance or whether there was an association between where Medicare beneficiaries lived and their satisfaction with services provided by carriers.

Table 1 is a contingency table of the joint observed frequency distribution of the two categorical variables "beneficiary location" and "satisfaction with carrier." For future reference, we have named the cells A-F.

Table 1: Observed Frequency Distribution

RESPONDENT	**SATISFIED**	**NEITHER SATISFIED NOR DISSATISFIED**	**DISSATISFIED**	**TOTAL**
Michigan *Observed*	Cell A 128 60%	Cell B 41 19%	Cell C 46 21%	215
National *Observed*	Cell D 206 75%	Cell E 32 12%	Cell F 38 14%	276
Total	334	73	84	491

In computing a Chi-Square test statistic, we compare observed frequencies with expected frequencies. Expected frequency is the number of observations in a cell that we would obtain if the two categorical variables were unrelated.

Computing expected frequencies for each cell is simply done by multiplying the row total by the column total and dividing by the sample size. For example, to find the expected frequency for Cell A (Michigan and Satisfied), you multiply 215 by 334 and then divide the total by 491. This results in an expected frequency of 146. Table 2

Table 2: Expected Frequency Distributions

RESPONDENT	**SATISFIED**	**NEITHER SATISFIED NOR DISSATISFIED**	**DISSATISFIED**	**TOTAL**
MICHIGAN *Observed* *Expected*	Cell A 128 *146*	Cell B 41 *32*	Cell C 46 *37*	215
NATIONAL *Observed* *Expected*	Cell D 206 *187*	Cell E 32 *41*	Cell F 38 *47*	276

shows observed and expected frequencies for each cell.

The next step in computing the Chi-Square test statistic involves subtracting the expected frequency from the observed frequency in each cell and squaring this value. After squaring the value, you divide by the expected frequency. Table 3 shows the computations for each cell.

Table 3: Computation of Chi-Square

Cell	O-E	$(O-E)^2$	$(O-E)^2$ ------ E
A	128-146	324	2.22
B	41-32	81	2.53
C	46-37	81	2.19
D	206-187	361	1.93
E	32-41	81	1.98
F	38-47	81	1.72
			12.57 = $\times^{2}$

The Chi-Square test statistic equals the sum of the values in the right hand column of Table 3. After you compute the Chi-Square test statistic, you must determine the degrees of freedom of your contingency table. You need the degrees of freedom to look up the critical values of the Chi-Square distribution at a specified significance level (alpha). The confidence level for the test is equal to (1-alpha) (i.e., the critical values associated with alpha = .05 corresponds to the 95 percent confidence level). In this example, the degrees of freedom equals two (2). This was computed by taking the number of rows minus one multiplied by the number of columns minus one: ((2-1)*(3-1) = 2). Tables of critical values for the Chi-Square distribution can be found in many statistical texts. A portion of the table of Chi-Square critical values is shown below.

Table 4:
Critical Values of the Chi-Square Distribution

df	a=.10	a=.05	a=.025	a=.010	a=.005
1	2.70554	3.84146	5.02389	6.63490	7.87944
2	4.60517	5.99147	7.37776	9.21034	10.5966
3	6.25139	7.81473	9.34840	11.3449	12.8381

a = alpha

The critical value of the Chi-Square distribution for 2 degrees of freedom at the .01

significance level is 9.21034. Since our Chi-Square test statistic was 12.57--a value greater than the critical value--we reject the null hypothesis that there is no relationship between where a beneficiary lives and their satisfaction with the carrier's service at the

.01 significance level. We believe that there appears to be some relationship between where a beneficiary lives and their satisfaction with their carrier's service.

Example 2 OEI Report: "Suppliers' Acquisition Costs for Albuterol Sulfate,"
OEI-03-94-00393

Using Chi-square Independence Test to determine if two categorical variables are statistically independent when one of the variables is respondent status (respondent/nonrespondent).

To determine suppliers' acquisition costs for the nebulizer drug albuterol sulfate, we asked suppliers to provide copies of invoices itemizing the prices they paid to acquire the drug. Suppliers were asked to return these invoices along with their responses to a self-administered questionnaire. We achieved an 86 percent response rate for the self-administered questionnaire. However, many suppliers did not provide copies of albuterol sulfate invoices. Nonresponse is a major source of survey error because it often introduces bias into sample data and can make a sample systematically different from the population from which it was drawn. With this in mind, we attempted (by telephone and letter) to secure the missing invoices. We created a binomial categorical data field in our survey database,"DDINVCS," to indicate invoice submission status. After our follow-up attempts, we achieved a response rate of only 47 percent for albuterol sulfate invoices.

With an invoice nonresponse rate of 53 percent, we had to ask ourselves the following questions:

1) Who were the suppliers that refused to submit an invoice?
2) Why would these suppliers refuse to submit an invoice?
3) What characteristics can tell us something about why they refused to submit an invoice?

We found that a few large-scale suppliers, with many albuterol sulfate claims in our sample, accounted for the invoice nonresponse. We needed to test for potential bias effects that these suppliers may have on our acquisition cost estimates. We reviewed National Supplier Clearinghouse (NSC) supplier profile data to learn about the suppliers in our sample (both respondents and nonrespondents) with respect to the size and scope of their business activities. Most of the variables maintained in the NSC database are quite incomplete. The most complete variable that we could use as a proxy for size of supplier business activity was the categorical variable "ASSETIND." This NSC field indicates if the supplier has assets over or under $100 million, or if the value is unknown. Therefore, we used the Chi-Square Independence Test to determine if supplier invoice submission status (respondent/nonrespondent) is statistically independent of the size of supplier assets.

In simple terms, the Chi-Square independence test would be used to answer whether the size of a supplier's business activity tells us something about their invoice submission status, and vice versa. We have two hypothesis from which to pursue the answer.

Null Hypothesis: Supplier invoice submission status and size of supplier assets are statistically **independent**.

Alternative Hypothesis: Supplier invoice submission status and size of supplier assets are statistically **dependent**.

SAS CODE

```
PROC FREQ DATA=A.AMY20715;
TABLES ASSETIND*DDINVCS/CHISQ;
RUN;
```

END SAS CODE

The SAS frequency procedure tells SAS to produce a contingency table for the two categorical variables "ASSETIND" and "DDINVCS" contained in the SAS data file "a.amy20715". The Chi-Square option tells SAS to produce Chi-Square test statistics and probabilities for the contingency table variables.

The following SAS Output displays the contingency table "ASSETIND" by "DDINVCS" and Chi-Square test statistics and probabilities.

Table 5:
ASSETIND BY DDINVCS

```
            DDINVCS

ASSETIND          Frequency
                  Percent   |
                  Row Pct   |
                  Col Pct   |        1|        2|  Total
                  ----------------------+--------+
                          1 |        5 |      95 |    100
                  Size      |     1.03 |   19.59 |  20.62
                  (<100,000)     5.00 |   95.00 |
                            |     2.09 |   38.62 |
                  ----------+---------+---------+
                          2 |      214 |     136 |    350
                  Size      |    44.12 |   28.04 |  72.16
                  (>100,000)    61.14 |   38.86 |
                            |    89.54 |   55.28 |
                  ----------+---------+---------+
                          3 |       20 |      15 |     35
                  Unknown   |     4.12 |    3.09 |   7.22
                            |    57.14 |   42.86 |
                            |     8.37 |    6.10 |
                  ----------+---------+---------+
                  Total          239       246      485
                                49.28     50.72   100.00


         STATISTICS FOR TABLE OF ASSETIND BY DDINVCS

      Statistic                        DF    Value      Prob
      --------------------------------------------------
      Chi-Square                        2     99.017 0.001
      Likelihood Ratio Chi-Square      2    117.072    0.001
      Mantel-Haenszel Chi-Square 1            69.929    0.001
      Phi Coefficient                          0.452
      Contingency Coefficient                  0.412
      Cramer's V                               0.452

      Sample Size = 485
```

Interpretation of SAS Output

When looking at the value of the Chi-Square test statistic, we want to answer the question, "Can this value be reasonably attributed to sampling error, or is it large enough to indicate that supplier invoice submission status and size of supplier assets are statistically dependent?" The larger the absolute value of the Chi-Square test statistic, the more likely that the two categorical variables are statistically dependent.

In this example, the value of the Chi-Square test statistic is 99.017 with 2 degrees of freedom. Further, there is only .001 probability that this Chi-Square test statistic occurred by chance/sampling error. Therefore, we reject the null hypothesis that supplier invoice submission status (respondent/nonrespondent) is statistically independent of the size of supplier assets at the .001 significance level.

Knowing something about supplier invoice submission status gives us information about the size of supplier assets, and vice versa. Suppliers possessing assets in excess of $100 million were supported with albuterol sulfate invoices. This percent was computed by dividing the frequency 5 (the number of suppliers with over $100 million who responded) by the row total of 100 (all suppliers over $100 million selected) and is given (in bold) as the row percent. In contrast, invoices were submitted for 61 percent of claims billed by suppliers with assets under $100 million. As given above, this percent was calculated by dividing 214 (the number of suppliers under $100 million who responded) by all 350 suppliers with under $100 million selected and is shown in bold.

We concluded, due to invoice nonresponse, our acquisition cost estimates are biased with respect to size of supplier business activity. We believe that large-scale suppliers with assets over $100 million may be able to use their market power to negotiate low costs for albuterol sulfate with drug manufacturers, wholesale outfits, and pharmacies. Therefore, our invoice cost calculations may actually overestimate average supplier acquisition costs.

***Example 3*: *Computing a Chi-Square test statistic using a Lotus-spreadsheet*.**

Another way to compute the Chi-Square test statistic is to use the Lotus-spreadsheet developed by TSS staff. This spreadsheet works with 2 x 2 tables and only requires you insert the frequency count for each cell and the column total. The table below illustrates the layout of the spreadsheet and the results you will get. If you do not have a copy of this spreadsheet, please contact the Technical Support Staff.

Table 6:
ASSETIND By DDINVCS
Using Spreadsheet

Spreadsheet to compute test statistic for Chi-Square and t-test						
	ASSETIND	DDINVCS	p	q	s.e.	
1	5	95	5.0%	95.0%	0.0218	100
2	214	136	61.1%	38.9%	0.0261	350
Total	229	234	49.5%	50.5%	0.0227	450
		t=	-9.90			
		Chi-sq=	98.151			

NOTE: Chi-Square and t statistic values correspond to particular confidence levels. The following are
standard Chi-Square & student's t distribution critical values.

CHI-SQUARE		t Statistic
3.84 or higher	= 95% confidence level.	1.95 = 95%
6.63 or higher	= 99% confidence level.	2.58 = 99%

CHAPTER 5

DETERMINING RELATIONSHIPS BETWEEN CONTINUOUS VARIABLES

Analysis Question: How do I determine if there is a difference between continuous variables?

ABSTRACT

Determining if differences exist between continuous variables is essential in analyzing data more completely. As discussed in the prior chapter, Chi-Square is used to test for differences when our variables are categorical. However, continuous variables require the use of t-tests. T-tests are used to compare key demographic variables such as the beneficiaries' age among respondents and non-respondents. This chapter explains the importance of using hypothesis testing, how to determine whether we should reject or fail to reject our null hypothesis, and what assumptions must be met before we can apply the t-test.

Introduction

There are several instances in our inspection work when it is necessary to compare the means of two continuous variables. A continuous variable is defined as a variable whose range (set of possible values) is an interval or a set of intervals on the real axis. Examples of continuous variables are age, allowed amounts, number of enrollees, etc.

In order to determine whether the means of two groups are significantly different, we develop what is called a "null hypothesis" (H_o). It states that two groups would have the same mean if the experiment were repeated a large number of times, and that differences in any one trial are attributable to random error. The alternative or test hypothesis (H_1) states that one particular mean will be greater than the other (a one-tailed test), or that the two means will be different, but we cannot say a priori which will be greater (a two-tailed test). A procedure called the t-test is used to determine the probability that the difference in the means of the two groups is due to chance.

We sometimes use a t-test to compare the ages of survey respondents and non-respondents to our surveys [such as those in Medicare beneficiary satisfaction surveys done by Region 4 (most recent report OEI-04-93-00150)]. A t-test can also be used to determine if responses to certain survey questions differ by age such as the survey of physicians interest in filing paperless claims (OEI-01-94-00230). Another example of the t-test is to compare the means of two samples, one before and one after a certain policy change has been implemented

Several t-tests were performed for the Region 6 inspection on the End-Stage Renal Dialysis (ESRD), "Know Your Number" brochure (OEI-06-95-00321), which randomly sampled 132 hemodialysis facilities about their experiences in distributing this brochure to their patients. One hypothesis explored was whether facilities varied in the way they rated the patient brochure based on how interested their patients were in finding out about adequacy information. Our alternative hypothesis said facility staff rating the brochure as excellent would also have more proactive patients; thus the means would not be equal. Our goal was to either reject or fail to reject the null hypothesis.

H_o: Mean of Group 1 (rated brochure excellent) = Mean of Group 2 (didn't rate excellent)

H_1: Mean of Group 1 does <u>not</u> equal Mean of Group 2

There are several assumptions that must be met before we can apply the t-test. First, the two groups must be independent. In our inspection work, we assume that respondents and nonrespondents are independent. Also, we will assume that the two groups are independent populations. In this case, one group consisted of facilities rating the dialysis brochure as excellent, while the other group consisted of facilities not rating it as excellent. The second assumption is that the theoretical distribution of sampling means is normally distributed. This can be achieved by selecting sample sizes of at least 30 from each group, although sample sizes of less than 30 can be normally distributed. In fact, as illustrated in an earlier chapter, we can test whether our data is normally distributed using PROC UNIVARIATE in SAS. For the example below, we assume the distributions are normally distributed. The two groups analyzed in this example contain 22 and 73 facilities each. Finally, the variances of the two groups should be approximately equal. This assumption is automatically checked by SAS each time the t-test is performed. Two sets of values will be given, one for equal variances and the other for unequal variances. We will explain later in this chapter how this appears in the output and how to interpret the results.

SAS CODE

```
PROC TTEST DATA=ESRD.III_ESRD;
CLASS Q15_EXC;
VAR Q7;
 RUN;
```

END SAS CODE

PROC T-TEST has two parts. The CLASS statement names the independent variable- the variable that identifies the two groups we are comparing. In this case, it is the group of dialysis facilities rating the brochure as excellent versus the group who did not. The variable on the VAR statement is the dependent variable. In the example that follows, the dependent variable is average percentage of patients in facilities interested in inadequate dialysis.

OUTPUT

```
Variable: Q7
Q15_EXC      N          Mean         Std Dev       Std Error        Minimum
Maximum
-------------------------------------------------------------------------------------------
   1.00         15     75.06666667     22.36855786      5.77553681
1.00000E+01     94.00000000
   2.00         57     52.33333333     28.20418746      3.73573589
0.00000E+00    100.00000000

Variances        T       DF    Prob>|T|
--------------------------------------
Unequal     3.3050     27.0      0.0027
Equal        2.8867     70.0      0.0052


For H0: Variances are equal, F' = 1.59    DF = (56,14)    Prob>F' = 0.3411
```

Interpretation of SAS Output

First, we need to examine the SAS results to determine which set of t statistics to look at, either those for equal or unequal variations. In this example, the SAS results show the variances of the two groups are equal. We fail to reject the null hypothesis shown above (H_o: Variances are equal) because Prob |T| =0.3411 is not significant. We look at the t statistic for equal variances instead. The t-statistic of 2.8867 is significant at the 99 percent confidence level since the Prob > |T| =.0052 (rounded to .01) and leads us to reject our original null hypothesis, concluding the means are not equal. If Prob > F had been significant at the 95 percent confidence level (equal to 0.05 or less), we would have used the t statistic corresponding to unequal variances (t=3.3050).

Reporting Results

OEI Finding in Report

Facilities rating the brochure as excellent for educating patients were significantly more likely to report more of their patients were interested in adequate dialysis. For example, facilities rating the brochure as excellent reported 75 percent of their patients were interested in adequate dialysis. In contrast, facilities who did not give the brochure an excellent rating reported only 52 percent of their patients were interested in adequate dialysis (see table below).

PERCENT OF PATIENTS REPORTED INTERESTED IN ADEQUATE DIALYSIS

By Facilities Rating the Brochure Excellent

Facility rated brochure as excellent	**Average percent of patients interested in adequate dialysis**
Yes	75%
No	52%

CHAPTER 6

ANALYZING DATA USING REGRESSION ANALYSIS

Analysis Question: How do I determine which key survey variables affect an outcome?

ABSTRACT

Once we perform a number of descriptive statistics, sometimes it is necessary to incorporate more advanced techniques. To estimate the relationship between a dependent variable and one or more Predictor (independent) variables linear or logistic regression must be used. Regression analysis also allows us to determine the direction of relationships. However, in addition to reading the material in this chapter, it is important you already have an understanding of basic statistics and work closely with a statistician in constructing and analyzing regression models for our reports. A statistics class in regression analysis would be extremely beneficial in becoming more aware of the caveats associated with using more advanced statistical methods.

Introduction

This chapter differs from others presented in this handbook. To fully understand regression analysis, it is recommended that you have taken at least one statistics course and have an understanding of terms such as parameters, estimates, distributions, mean and variance of a random variable, covariance between two variables, and simple hypothesis testing involving one- and two-sided t-tests and the F test. Semester long college courses are devoted to using regression techniques. As a result, it is impossible to try and present a step-by-step approach on using SAS to perform regression analysis in this brief chapter of the handbook. In practice, some diagnostic tests of your regression model will likely be needed, and the model must also fit a set of stringent assumptions. If you have not been trained in regression analysis, further reading and consultation with a statistician or econometrician will be necessary to conduct this type of analysis. However, since regression analysis is a powerful and useful tool, we do want to introduce it and show the SAS code and interpretation of the results using an OEI example to demonstrate how we have been able to use it in the past.

Statistical Associations

In analyzing data, we often want to determine whether there is an association between key inspection variables. For example, what factors predict a person's annual income (dependent variable)? Is there a relationship between a person's gender and their income? Using a simple t-test we could determine the level of statistical significance between these two variables. However, could other factors (known as independent variables) such as a person's job experience, age, IQ, race, parent's income, or educational level also influence one's income (known as the dependent or outcome variable)?

Relying on statistical associations, using statistics such as t-tests or Chi-Square, only allow us to measure the relationship between variables and not whether the relationship is positive or negative. Nor do they allow us to measure the unique effect an individual variable has on a dependent variable when numerous variables are included in the model. We are also unable to determine which variable(s) have the greatest impact on the outcome variable. Additionally, tests for associations are sensitive to sample size so a given relationship is more likely to be significant in a large sample.

To accurately analyze which independent variables influence a particular outcome variable, such as one's income, other variables must be considered and controlled. In order to do this and also avoid the adverse consequences of relying on tests of association, it becomes necessary to move into regression analysis.

Elaboration of Statistical Associations Using Regression Models

Regression analysis permits us to further explore what happens to bivariate (two-way) associations once new, additional variables are taken into account. Only through a multivariate regression model will we be able to fully test the simultaneous impact numerous variables have on a particular outcome. Specifically, for each independent variable, the slope is calculated for a line which most closely fits the relationship between the independent variable and the dependent variable. For example, a one year increase in job experience leads to a $1.00 per hour increase in wages. Each independent variable will have an estimated coefficient (slope) that measures the effect of the variable, after controlling for the other variables in the equation. This means you can measure the unique effect of an independent variable on the dependent variable.

However, relying just on the correlation between one independent and the dependent variable is not sufficient to fully answer your questions about such relationships. In fact, they often cause you to get results which can be extremely misleading and even damaging. Many times an original relationship is explained away, or the significance of the relationship reduced, by a third variable. For example, suppose you want to determine how key variables help predict future income. After running a simple t-test you determine that there is a significant relationship between one's gender and income, suggesting gender discrimination. You also find an association between the years of a person's work experience and income amount. After estimating a linear regression equation, which includes this new variable, the significance of the gender variable on one's income decreases dramatically. The factor which has the greatest effect on income is not gender, but really years of work experience. This suggests that gender discrimination in the workforce might not be as dramatic as you had originally found before you controlled for a person's years of work experience. Without this further elaboration of the relationships in the data, misguided policy recommendations could have resulted.

Technical Aspects of Regression Analysis

Regression analysis uses one or more independent variables to explain an outcome, or dependent variable. There are several types of regression techniques used to analyze data; two major types are explained in this chapter, linear and logistic regression models. An analysis involving a continuous dependent variable, such as annual salary, generally requires the use of a *linear* regression model. An analysis using individual-level data and having a dichotomous (two possible values: either a one or zero) dependent variable generally requires the use of a *logistic* regression model.

The Model Building Process

The following outline presents some of the steps involved in conducting a regression analysis.

1. Planning Stage

   - Define the research question. Will one variable have a causal effect on another? For example, does gender affect salary?

   - Develop a hypothesis. A hypothesis is a statement about predicted relationships among events or variables. For example,
     H_0 = Gender has no effect on salary.
     H_1 = Gender does effect salary.

- Determine other control variables. For example, which control or intervening variables would we expect to influence salary? Which variables are available?

  Possible Control or Independent Variables for a model predicting income include:

  a. Educational major
  b. Educational level
  c. Gender
  d. Years work experience
  e. Age
  f. Marital Status
  g. Race
  h. Work Skills

- Decide amount of acceptable variation. Depending on the type of study you are conducting, this acceptable variation will differ. For example, a medical or economic study might require a higher degree of explanatory power of the independent variables compared to a study measuring human behavior.

- Run correlation matrix (covariance matrix). Test the relationships between your independent variables. Are you measuring the same thing twice? For example, one's work skills and educational major might be highly correlated (0.80 level or above). If they are correlated, you would only use one of these variables in your regression model.

2. Developing a Regression Model

   - Plot variables, equation, residuals. The dependent variable should be plotted against the various independent variables.

   - Refine regression model. Are there additional variables that are needed or should some be removed from the model?

   - Consult experts for criticisms.

   - Plot any new variables, re-run correlation matrix, and interpret residuals. Residuals from any fitted equations should be plotted against any new variables.

3. Evaluation of Regression Model

   - What do your results mean? What do your results say about your hypothesis? Is the model plausible and usable? Does it make sense? Consult other researchers in the field you are studying.

- Is your regression model adequate? Run diagnostic tests. Is there systematic lack of fit? By examining the residuals, it is possible to recognize a problem such as the omission of key variables.

- Look at R-square (Goodness-of-fit) for standard linear regression model or the adjusted R-square for logistic regression models.

- What does your model say about your control variables? Were they statistically significant predictors of your outcome?

- Is your model applicable over time?

Further Developing of a Regression Model

A regression model specifies the equation for the line(s) which most nearly fit the relationships between the independent and dependent variables. Ideally, the equation will include all independent variables which have a significant unique association with the dependent variable. The goal in building either linear or logistic regression models is to include hypothesis testing and theoretical research in your regression models. For example, what does the literature already say about your particular topic? Unlike much academic research, our work topics are often exploratory in nature and have not been thoroughly studied by other researchers. If research exists on your topic, your regression model should be based on this research and should help guide you in developing your survey instrument or the data elements you request for your dataset. For example, if the research shows that age also affects annual income then you would want to be sure you have a variable measuring age.

You will always want to include key demographic variables about both individuals in your sample, (i.e., race, educational level, gender) or groups in your sample (i.e., region of the country, facility size, or ownership type). If no literature exists about your inspection topic, exploratory research is done to develop a model that can be used as a baseline for future research. It is extremely important that all variables incorporated in the final regression model be based on your theoretical research or plausible explanations for the relationships that are identified.

Testing for Correlation between Variables

To test whether or not there are any relationships between independent variables in your model, it is important to examine pairwise correlations among all variables. By running a correlation matrix using SAS, you will be able to determine variables which variables are related to each other. Perfect correlation would result in a correlation coefficient of 1.0, meaning you have identical responses for every observation for a pair of variables.

An example of correlated variables in a prior OEI study involved two survey variables that questioned beneficiaries about their reading and writing skills. Based on a correlation coefficient of 0.90 for this pair of variables, we found that beneficiaries in our dataset reporting they had reading difficulties also had writing difficulties. In this case, we only included one of the two variables in our regression model. Including variables that are highly correlated (usually 0.80 or above) with each other will reduce the efficiency of the variable coefficients in your model.

SAS CODE

```
PROC CORR  DATA=SSA;
MODEL HARD= READING  WRITING......
RUN;
```

END SAS CODE

Linear Regression Models: An OEI Example

Linear regression analysis is used when the dependent variable is continuous and numeric, such as annual salary. Provided the dependent variable is continuous, several or all of the independent variables may be binary (one or zero) or categorical variables. In the OEI report "Medicare Risk HMO Performance Indicators" (OEI-06-91-00734), several examples of linear regression models are found. To help identify potential independent variables for inclusion in the regression models, we relied on the limited literature available on Medicare risk HMO disenrollment rates and key associations discovered in another OEI report, Medicare Perspectives of Medicare Risk HMOs (OEI-06-91-00730).

We hypothesized that HMOs with a higher percentage of unhappy beneficiaries would lead to higher future beneficiary disenrollment in our sampled HMOs. However, we did not know which key variables from our survey data of Medicare risk HMO beneficiaries would be the best predictors of our HMOs future disenrollment rates. By using a regression model, we wanted to identify the significant predictor variables that we could recommend for researchers and the Health Care Financing Administration (HCFA) to focus on when evaluating

Medicare risk HMOs.

Independent Variables

Certain demographic variables about HMOs were used as control variables in the model. Many of these variables were categorical variables and needed to be dichotomized (0 or 1) for each category prior to using them in our regression model. To avoid a major mistake in dealing with these variables, one of the categories must serve as the referent category. For example, for the variable measuring the HMO model type (Staff, Group, or IPA), we only included two of the three dichotomized variables in our regression model. Both the staff and group models were included in the model, while the IPA model served as the referent category. The same method was used in handling variables measuring the region of country in which HMOs were located. For example, South, Northeast, and Midwest were included in the model, while West served as the referent category. Other demographic variables included in the model were HMO size, HMO profit status, competitive status, and the percentage of beneficiaries from each HMO who received their care through fee-for-service prior to joining the HMO. Table 1 provides a complete description of these variables.

Table 1: INDEPENDENT VARIABLE DESCRIPTIONS
Key Demographic Variables

Variable	Description
STAFF	1 = Staff HMO, 0 = IPA HMO
GROUP	1 = Group HMO, 0 = IPA HMO
SIZE	# of Medicare Risk Beneficiaries enrolled in HMO
SOUTH	1 = HMO located in South, 0 = HMO located in West
NEAST	1 = HMO located in Northeast, 0 = HMO located in West
MIDWEST	1 = HMO located in Midwest, 0 = HMO located in West
FFS	Percent of beneficiaries receiving prior care through fee-for-service
PROFIT	1 = HMO is for profit, 0 = non-profit HMO
COMPETE	1 = HMO is in competitive area, 0 = HMO is in non-competitive area

After controlling for all the demographic variables, we wanted to see which of the key variables in Table 2 were significant predictors of HMOs disenrollment rates. The non-demographic independent variables included in the model were based on prior literature and key associations discovered in our prior report.

Table 2: INDEPENDENT VARIABLE DESCRIPTIONS
Key Variables of Interest (Non-demographic)

Variable Name	Description
POOR_SER	Percent of beneficiaries with HMO doctor not providing Medicare services, hospital care, or referral to specialist care
SICK	Percent of beneficiary-reported serious health problems
Q_HEALTH	Percent of beneficiaries asked questions about health problems, at time they applied for HMO membership
HOUR_WT	Percent of beneficiaries usually waiting an hour or more before seeing their primary HMO doctors
COMPLAINT	Percent of beneficiaries reporting their primary HMO doctor did not take their complaints seriously
OWN_SERV	Percent of beneficiaries in last year, getting Medicare covered services on their own without primary HMO doctor or HMO first approving
HMO_PRIOR	Percent of beneficiaries reporting their HMO most concerned with holding down the cost of medical care

SAS CODE

```
PROC REG DATA=HMO.ENRHMO;
MODEL NADM_92= PROFIT POOR_SER  COMPLAINT
STAFF SOUTH
SIZE GROUP  FFS COMPETE  SICK NEAST MIDWEST
HOUR_WT Q_HEALTH
OWN_SERV  HMO_PRIOR /selection= FORWARD
SLENTRY=0.05;
```

```
        RUN;
```

END SAS CODE

The PROC REG procedure invokes the SAS regression for the specified dataset. The MODEL statement is in the form: dependent variable=independent variable(s). The selection=FORWARD indicates that the method for variable selection is the FORWARD method. There are several types of methods for variable selection. A discussion of the merits of each type of method (forward, backward, and stepwise) may be found in a regression textbook. The FORWARD method begins by finding the variable that produces the optimum one-variable model and tries to improve on the model by adding variables one by one until no variable considered for addition to the model provides a reduction in sum of squares considered statistically significant at a level specified by the user. This level is given by the SLENTRY option. In the example above, a SLENTRY of .05 was chosen meaning that a variable has to be significant at the 95 percent confidence level for providing a reduction in the sum of squares.

Linear Regression Results

The following table shows the overall regression results of our reduced models used in the report. Some key information from the SAS output includes the non-standardized regression coefficient (parameter estimate), intercept, standard error, t-value for all variables, significance level (probability $> |t|$), and the R^2.

```
        Model: MODEL1
        Dependent Variable: NADM_92

                                   Analysis of Variance

                                  Sum of         Mean
           Source         DF      Squares       Square      F Value       Prob>F

           Model           3    961.31976    320.43992        8.816       0.0002
           Error          35   1272.20227     36.34864
           C Total        38   2233.52204

               Root MSE       6.02898     R-square       0.4304
               Dep Mean       9.75795     Adj R-sq       0.3816
               C.V.          61.78535

                             Parameter Estimates

                                Parameter      Standard       T for H0:
          Variable           DF             Estimate         Error
Parameter=0    Prob > |T|

          INTERCEP           1     -5.193875     3.36713875       -1.543
        0.1319
          PROFIT             1      0.048672     0.02314383        2.103
        0.0427
          POOR_SER           1      0.684686     0.23699629        2.889
        0.0066
          COMPLAINT          1      0.515142     0.20921318        2.462
        0.0189
```

Intercept and Regression Coefficients

The *intercept* is a constant which is added to the regression model and is based on the "Y" value when all independent variables are set to zero. In the regression equation shown below, the intercept is -5.19. Because the intercept is negative, we subtract it from the value predicted by the regression equation.

The *nonstandardized regression coefficient* indicates the change in "Y" associated with a one-unit change in the independent variable, while holding the other variables constant.

The nonstandardized coefficient represents the “raw score” form. Because independent variables have their own unit of measurement, with different means and standard deviations, the nonstandardized coefficients in your model cannot be assessed for their relative importance between variables.

Regression Equation

Based on the SAS output shown previously:

Y’ = -5.19 (intercept) + 0.049 (profit) + 0.685 (poor_ser) + 0.515 (complaint)

In prior OEI studies, we have usually shown the nonstandardized regression coefficient. However, you may choose to use the standardized regression coefficient (beta weights) instead. Although not shown in our SAS output, beta weights are the coefficients that would be obtained if all variables were standardized. Thus, giving them the same unit of measurement (mean of zero and standard deviation of 1). By standardizing regression coefficients, you can reflect the relative importance of the various variables in your model.

Standard Errors

The standard error for each coefficient estimate is a measure of sampling error or the errors in our estimates due to random fluctuations in our sample. The smaller the standard error, the better the sample statistic estimates the population parameter. The parameter estimate divided by the standard error gives us the t-value for each independent variable.

t-Value

The t-value for each variable tests for the significance of each independent variable’s effect on the dependent variable. The last column on the SAS output gives the probability of the t-value. The t-values and the associated probabilities (probability > |t|) test the null hypothesis that, in the population, the regression coefficient is equal to zero and answers the question, “If the true slope or intercept were zero, what would the probability be of obtaining, by chance alone, a t-value as large or larger than the one actually obtained?”

Because the example shown on the prior page used a forward selection method and SLENTRY=0.05, only coefficients with a *p*-value of .05 are listed. The independent variable POOR_SER is significant at the .01 level since the corresponding *p*-value =

.0066. All three independent variables listed in the SAS output allow us to reject the null hypothesis and conclude these estimated parameters are significantly different from zero.

R^2 and Adjusted R^2

The R^2 indicates the percent of variance accounted for by the linear combination of independent variables. In the following SAS output, a R^2 of .43 indicates the linear combination of the three variables accounts for about 43 percent of the variance in future HMO disenrollment rates. Also listed, is the adjusted R^2 which accounts for degrees of freedom: N-1 or the number of observations in your sample minus 1. The adjusted R^2 is considered a more conservative measure because it more closely approximates the population's true value. In this example, the adjusted R^2 is .38.

Interpretation of SAS Output

Each parameter estimate measures a one unit change in the associated independent variable and predicts an increase (or decrease if the sign of the parameter is negative) equal to the parameter value, holding all other independent variables constant. For example, in the SAS output table, the estimated coefficient for the COMPLAINT variable is 0.51. This means that for each one percent increase in the proportion of beneficiaries who felt that their doctor didn't take their health complaints seriously there is a 0.51 percent increase in the disenrollment rate for an HMO, holding the POOR_SER and PROFIT variables constant. The probability that this parameter is at least 0.51 is 0.0189, more than meeting a .05 significance level criteria.

Discussion of Results: OEI Finding

HMOs with higher disenrollment rates had more enrollees who reported service access problems.

Enrollees who said they experienced poor service, whose complaints were not taken seriously, and who were in for-profit HMOs, were more likely to come from HMOs with higher disenrollment rates. These three factors helped explain much of the variation in our HMOs' Calendar Year 1992-93 disenrollment rates, even after controlling for such structural characteristics as HMO type and enrollment size (see Table 3).

Table 3: DEPENDENT VARIABLE = Annualized adjusted disenrollment rates for '92 & '93[*]
N = 39 HMOs, enrollees only[**]
R^2 = .43 & Adjusted R^2 = .38

Variable	Parameter Estimate	Standard Error	t-Value	Probability > |t|
INTERCEPT	-5.19	3.37	-1.54	.1319
POOR_SER	0.68	0.24	2.89	.0066
COMPLAINT	0.51	0.21	2.46	.0189
PROFIT	0.05	0.02	2.10	.0427

[*] The adjusted disenrollment rate excludes the administrative disenrollees (based on a formula created by our survey data, see the report for further explanation)

[**] Excludes 6 HMOs; 2 dropped their risk HMO contract, 3 had a low number of disenrolled respondents, & 1 split into two risk HMOs, transferring many beneficiaries into a new HMO.

Logistic Regression Analysis: An OEI Example

Logistic regression analysis is used when the dependent variable is dichotomous or binary (one or zero). The "Medicare Risk HMO Performance Indicators" (OEI-06-91-00734) report highlights a logistic regression model. Because all logistic regression analysis must be done using individual-level data as the unit of analysis (i.e., a beneficiary or a state); we used key information about beneficiaries' experience with their HMO. Our dependent variable measured whether or not beneficiaries disenrolled from their HMO. With dichotomous variables, we do not want to use Ordinary Least Squares (OLS) regression analysis, which is used for linear regression analysis. This is because the resulting estimates are not the closest fit to the data and OLS will predict outcomes less than zero or greater than one. The technique we use instead is called logistic regression because the probability distribution for the dependent variable has a logistic, rather than a normal, distribution. The mathematical technique for logistic regression is called Maximum Likelihood Estimation rather than the OLS. For a brief description of these techniques refer to the glossary. A listwise deletion was used in this analysis so only beneficiaries who answered all of the questions were included in the regression model. Table 4 includes a complete list of all variables used in this model.

Table 4: VARIABLE DESCRIPTIONS
DEPENDENT VARIABLE: DISENROLLEE

Variable Name	Description
SEX	1 = Beneficiary was male, 0 = female
DISABLED	1 = Disabled, 0 = Aged beneficiary
AGE	Beneficiaries' age
COMP_AREA	1 = Beneficiary lived in a non-competitive area, 0 = Beneficiary lived in competitive area
PROP_USE	1 = Low/Medium use of medical services by beneficiary, 0 = High use of medical services
SICK	1 = Beneficiary was rated as very sick, 0 = Beneficiary was not rated as moderately sick or not sick
PRIOR_HMO	1 = Beneficiary came previously from a HMO, 0 = Beneficiary came previously from fee-for-service
HMO_PRIOR	1 = Beneficiary reported their HMO was most concerned with holding down the cost of their medical care, 0 = Beneficiary reported their HMO was most concerned with providing the best care possible

The logistic regression model illustrated here measures the effect of beneficiaries' negative HMO experiences on the likelihood of disenrollment, even after controlling for a number of demographic characteristics, such as age, sex, and health status.

SAS CODE

```
PROC LOGISTIC DATA=HMO1.SMHMO6;
MODEL STATUS = SEX DISABLED AGE COMP_AREA
PROP_USE SICK PRIOR_HMO HMO_PRIO;
RUN;
```

END SAS CODE

Logistic Regression Results

The variables created for this analysis are binary (except age), characterizing beneficiaries with either a one (1) if the characteristic existed or a zero (0) if otherwise (see Table 4 above). For this regression model, beneficiaries or events coded as zero represent the referent (the one which the other category will be compared to) category. The dependent variable estimated in this model, disenrollment status, measured the change from the referent status (enrolled) produced by each independent variable.

We used SAS to run this logistic regression model. The model allows us to estimate the probability of a beneficiary disenrolling from their HMO (ϱ) or remaining enrolled ($1-\varrho$), based on the linear combination of independent variables.

That is, $\ln(\varrho/1-\varrho) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$

where β_i is the coefficient estimated by the equation, X_i is the value of the independent variable, and k is the number of independent variables in the equation, and $\ln(\varrho/1-\varrho)$ is the log of the odds ratio of the dependent variable having a value equal to one (disenrollees).

```
Data Set: HMO1.SMHMO6
    Response Variable: DISENR                       The LOGISTIC Procedure
    Response Levels: 2
    Number of Observations: 674
    Link Function: Logit
                                                     Response Profile
                                                            Ordered
                                                    Value  DISENR     Count

                                              Enrollee             1      0       505
                                                    Disenrollee          2      1
                                                    169

                                          Testing Global Null Hypothesis: BETA=0
                                                    Intercept
                                    Intercept             and
                  Criterion          Only          Covariates              Chi-Square for
                                                                           Covariates

                   AIC                   761.124               713.134                 .
                   SC                    765.638               776.319                 .
                  -2 LOG L               759.124               685.134          73.990 with
                                                                               13 DF
                                                                               (p=0.0001)

                   Score                      .                    .            80.424 with 13 DF
                                                    (p=0.0001)

                                     Analysis of Maximum Likelihood Estimates
                        Parameter     Standard        Wald                   Pr >        Standardized
        Odds
 Variable         DF       Estimate         Error              Chi-Square    Chi-Square
Estimate          Ratio
 INTERCPT     1      -7.5035          1.8258          16.8892         0.0001
 SEX            1      0.4607          0.1915          5.7847          0.0162
0.126435      1.585
 DISABLED     1        -0.0604         0.4683           0.0166         0.8974
-0.007588             0.941
 AGE             1       0.0172          0.0164           1.0990         0.2945
0.066053      1.017
 COMP_AREA1      0.0782          0.1927           0.1646         0.6850
0.021552      1.081
 PROP_USE    1        -0.0627         0.2515           0.0621         0.8032
-0.012918             0.939
 SICK            1       0.0790          0.3651           0.0469             0.8286
```

Interpretation

Using the logistic regression model, interpretation of the coefficients, β_i can be translated easier using the exponential of the coefficient, known as the odds ratio. The estimated odds ratio reflects the extent to which the independent variable increases or decreases the odds that a beneficiary disenrolled from their HMO. An odds ratio of 2.4 for a particular independent variable, such as HMO_PRIO, is translated as: the beneficiary is 2.4 times more likely to have disenrolled, if they said their health got worse as a result of the care received by their HMO. Significance for each variable is determined by use of a Chi-Square distribution rather than the t-distribution used in linear regression.

Highlighted variables are statistically significant at the .10 confidence level or better, as found under the column "Pr > Chi-square." It is common to refer to variables being significant at the .01, .05 or .10 confidence levels. For example, the variable HMO_PRIO is significant at beyond the 0.05 confidence level (p=.0001). For this variable, it is true that 95 out of 100 times this parameter estimate will approximate the true value in the population.

The standardized coefficient provides a way to rank the relative importance of the variables according to their effect on beneficiary disenrollment from their HMO. Higher absolute values imply a greater effect. For example, the variable HMO_PRIO has the highest absolute value in this model.

There is also a way to determine how well the model fits the data by determining the concordant or C value for the equation. This represents the percentage of times for which the model accurately predicts the observed outcome (enrollment/disenrollment). For this model, the C value is .69. Values greater than 0.5 indicate a fit and values close to 1.0 are the ideal.

Discussion of the Results

In this example, variables influencing beneficiaries to disenroll, after adjusting for everything else include 1) beneficiaries' HMO prioritizing holding down the cost of medical care and 2) being a male beneficiary.

Additional Reference Books

Freund, Rudolf and Ramon Littell. 1992. SAS System for Regression: Second edition. SAS Institute Inc., Cary, N.C.

Hatcher, Larry and Edward Stepanski. 1994. A Step-by-Step Approach to Using the SAS System for Univariate and Multivariate Statistics. SAS Institute Inc., Cary, N.C.

Khattree, Ravindra and Dayanand Naik. 1995. Applied Multivariate Statistics with SAS Software. SAS Institute Inc., Cary, N.C.

Menard, Scott. 1995. Applied Logistic Regression Analysis. Sage Publications. 1995.

PART II

DATA MANIPULATION AND SAMPLING

CHAPTER 7

SELECTING SIMPLE AND STRATIFIED RANDOM SAMPLES

Analysis Question: How do I select a simple or stratified random sample?

ABSTRACT

Nearly every inspection requires selecting a sample from a population. One of the most common methods used is systematic sampling, which approximates simple or stratified random sampling. To select a systematic sample, we start with a population file that is in random order, and then develop a selection criteria for generating the sample, such as every 100th beneficiary. This chapter describes the process for selecting a simple and stratified random sample, includes an in-depth explanation of the SAS program language, and illustrates how to a use PROC FREQ to provide population counts for each strata.

Introduction

Drawing a sample is a common occurrence in OEI. Systematic random sampling was used to select the sample given in the examples below. To select a sample, we take a unit at random from the first K units and every Kth unit thereafter when K = N/n. For example, if K is 15 and the first unit drawn is the 13th on the file, the subsequent units are numbers 28,43,58, and so on. Systematic sampling is used by OEI, because it simplifies the sample selection process. Also, it is essentially equivalent to simple random sampling (selecting random numbers from a table or generation by a computer) under certain conditions. The main condition we must first meet and understand is that the population we sample from is in “random” order. In other words, if a file is in HICN order, we assume the item being measured has no relation to the HICN of the individual.

Standard Process for Selecting a Sample Using SAS

First start with a master database; for example, all hospital inpatient claims in 1995. Apply a pre-determined set of criteria to the database creating a subset file; for example, all patients admitted between June 1 and December 31 with a diagnosis of stroke. The sample would be selected from this subset, which is our sampling frame.

The SAS programs for selecting a sample can be used on either a PC or mainframe computer. Usually sample selection is performed on the mainframe due to the file's size and then the sample file is downloaded to a PC for analysis and manipulation. Before doing the actual procedure, you need to know the total number of records in your subset file and the number of records wanted in the sample. Use a PROC FREQ, PROC SUMMARY, or PROC UNIVARIATE to determine the number of records in the subset file (population or sampling frame).

SAS CODE

```
IF _N_= 1  THEN DO;
F=_FREQ_/SAMPLE;
G=UNIFORM(0)*F;
N=1;
END;
RETAIN N F G;
IF N GE G THEN DO;
OUTPUT;
N=N-F+1;
END;
.        ELSE N+1;
```

END SAS CODE

The program works as follows:

1. A DO loop is set up within each group to be sampled.
2. A skip interval (F) is computed using the population and sample size for each group.
3. A random starting point (G) is selected, using the SAS function UNIFORM, between one and the skip interval.
4. The process begins with the first observation on the file.
5. A comparison is made between the observation number and the random start.

6. If the observation number is greater than or equal to the random starting point, it is selected for the sample. Otherwise, the file is advanced to the next record.
7. Once a record is chosen, the skip interval is subtracted, and 1 is added to begin the DO loop comparing N to G all over again. This process continues until the entire file has been subjected to sampling.

In line 2 of the program above, _FREQ_ is a generated variable in a PROC SUMMARY. It is a count of the number of records. In SAS, _N_ is another way to look at or use the observation number. It is kept internally in SAS, does not show as a variable on a printout, and changes as the record is put into different files. In the following example, _N_ (see line 3) is used to get to the first record in the file to be sampled.

Example 1

In the example below, I wanted 840 records in my sample. The population sampled had 304,435 total records. This program illustrates a simple random sample without stratification or grouping of the data.

SAS CODE

```
DATA OUT.S1 (DROP= N F G);
SET IN.S1;
IF _N_ =1 THEN DO;
F=304435/840;
G=UNIFORM(0)*F;
N=1;
END;
RETAIN N F G;
IF N GE G THEN DO;
OUTPUT;
N=N-F+1;
END;
ELSE N+1;
RUN;
```

END SAS CODE

Example 2

This example shows how to set up your strata and then select a sample from within each strata. This program is best used when you only have a few strata.

SAS CODE

```
DATA STRAT1 STRAT2;
SET IN.S1;
IF STATE IN (‘CA’ ‘FL’ ‘TX’ ‘IL’ ‘NY’) THEN
        OUTPUT STRAT1;
ELSE OUTPUT STRAT2;
DATA OUT1;
SET STRAT1;
IF _N_ =1 THEN DO;
F=1572/750;
G=UNIFORM(0)*F;
N=1;
END;
RETAIN N F G;
IF N GE G THEN DO;
OUTPUT;
N=N-F+1;
END;
ELSE N+1;
DATA OUT2;
SET STRAT2;
IF _N_ =1 THEN DO;
F=3248/250;
G=UNIFORM(0)*F;
N=1;
END;
RETAIN N F G;
IF N GE G THEN DO;
OUTPUT;
N=N-F+1;
END;
ELSE N+1;
DATA OUT.S1;
SET OUT1 OUT2;
RUN;
```

END SAS CODE

This program was written for a sample divided into only two strata (Operation Restore Trust (ORT) States and non-ORT States). A previous PROC FREQ on STATE provided the strata population counts. Lines 1 to 4 create two strata, ORT States versus all other States . The variable STATE is used to group the data. Lines 5 and 6 are accessing only the records from the ORT State's strata. We selected 750 units from the ORT State's strata. Lines 18 and 19 are using only records from the second strata (all other States) from which we selected 250 records. Lines 31 and 32 join the 750 records selected from the first strata and the 250 from the second into one newly created file (OUT.S2) containing all the sampled records.

Example 3

This third example also illustrates how to set up your strata and then select a sample. However, in contrast to when you have few strata, this SAS program is a more effective example for using when you have many strata or when the population file is very large.

In this example, we use GHPNO as the stratification variable. We wanted 550 cases sampled if GHPNO equaled 'H1036' and 133 cases sampled for **each** of the other values of GHPNO. Lines 10 through 30 perform the same task as lines 5 through 17 in the previous example. This is the process of picking a random start, calculating a skip interval, and selecting the appropriate observation. It appears differently, because words have been substituted for the letters N, F, G. We purposely included this example to show how the sampling program can be integrated into other programs. This program also uses a PROC FREQ to provide the population counts dynamically (i.e., instead of coding the numbers into the calculation, a variable name is used).

SAS CODE

```
PROC FREQ DATA=IN;
TABLES GHPNO /OUT=NUMCNT NOPRINT;
PROC SORT DATA=NUMCNT(KEEP=GHPNO COUNT);
BY GHPNO;
PROC SORT DATA=IN OUT=INSRT; BY GHPNO;
DATA UNIV;
MERGE INSRT (IN=MASTER)  NUMCNT (IN=GHP);
BY GHPNO;
IF MASTER AND GHP;
DATA OUT.S1(DROP=REC INTERVAL PICK OFFSET COUNT);
SET UNIV;
BY GHPNO;
RETAIN REC INTERVAL PICK;
IF FIRST.GHPNO THEN DO;
IF GHPNO=’H1036' THEN DO;
INTERVAL=COUNT/550;
OFFSET=UNIFORM(0)*(COUNT/550);
END;
ELSE DO;
INTERVAL=COUNT/133;
OFFSET=UNIFORM(0)*(COUNT/133);
END;
PICK=OFFSET;
REC=1;
END;
IF REC GE PICK THEN DO;
OUTPUT;
PICK+INTERVAL;
END;
REC+1;      RUN;
```

END SAS CODE

DATA EXAMPLE for program above.

Data File Grouped by "GHPNO"

```
     GHPNO          COUNT
     PERCENT

     H0033           1456
25.0
     H1036           2256
40.0
     H1222            365
5.0
     H1234           1133
20.0
     H4786            100
2.0
     H5432            200
3.5
     H6543            325
4.5
```

Result of MERGE:

Merged Master & GHP Files -----------

GHPNO	COUNT	PERCENT	BENENUM	ZIP
H0033	1456	25.0	2546	21044
H0033	1456	25.0	2546	21044
H0033	1456	25.0	2546	21044
H1036	2256	40.0	6543	21007
H1036	2256	40.0	6543	21007
H1036	2256	40.0	6543	21007
H1036	2256	40.0	6543	21007
H1036	2256	40.0	6543	21007
H1036	2256	40.0	6543	21007
H1222	365	5.0	1234	21207
H1222	365	5.0	1234	21207
H1222	365	5.0	1234	21207
H1222	365	5.0	1234	21207
H1234	1133	20.0	4325	21113
H4786	100	2.0	600	21114
H4786	100	2.0	600	21114
H4786	100	2.0	600	21114

H5432	200	3.5	800	20708
H6543	325	4.5	1256	20800

Results of sample selection:

The PROC FREQ (lines 1 and 2) provides the population counts for each value of GHPNO. The /OUT= option on line 2 creates a dataset of the results of the frequency distribution. The two PROC SORT statements are used to ensure both files are in the correct order for the MERGE. In lines 7 to 9, MERGE appends COUNT (population total) to each record in the universe based on GHPNO's value. On line 14, the FIRST. statement creates a logical variable which is used to determine the first record, last record, and in-between records in a sorted file.

Selection of Stratified Samples

GHPNO	POP. COUNT	# SELECTED
H0033	1456	133
H1036	2256	550
H1222	365	133
H1234	1133	133
H4786	100	100
H5432	200	133
H6543	325	133
TOTALS	5835	1315

These next two charts show, in general, what a set of data looks like before and after applying FIRST. and LAST statements. In this case, HICN is the sorted variable.

Sorted data **before** applying FIRST. or LAST.

Sorted Data Before

HICN	ALLW	TOTSVCS
123aa	23.40	4
123aa	100.00	1
123aa	10.50	1
136bb	15.20	3
136bb	200.00	6
175aa	20.20	2

Sorted data **after** applying FIRST. or LAST.

Sorted Data After

```
HICN   ALLW   TOTSVCS  FIRST
LAST

123aa           23.40           4
   1    0
123aa          100.00           1
   0    0
123aa           10.50           1
   0    1
136bb           15.20           3
   1    0
136bb          200.00           6
                  0     1
175aa           20.20           2
  1     1
```

When querying the FIRST and LAST variables, you are looking at whether they are set as true (1) or false (0). To use this logic, the data needs to be sorted by the variable being tested (see GHPNO in the program on the second page of this chapter). Also, SAS has to be told the data is sorted by that variable. The code for this notification follows:

SAS CODE

```
DATA OUT;
SET IN;
BY SRTVAR;

/*To use SRTVAR it would look like this*/

IF LAST.SRTVAR THEN CUMTOT + ALLW;
IF FIRST.SRTVAR THEN DO;
CUMTOT=0;
END;

/* is a comment in SAS*/
```

END SAS CODE

CHAPTER 8

ANALYZING BENEFICIARY PAYMENT HISTORY

Analysis Question: How do I look at a billing history for a sample of beneficiaries?

ABSTRACT

We often need to analyze the billing history of sample beneficiaries to see how their claims fit into their medical history from a logical point of view. It further assures us whether health care providers are billing for services provided. This chapter describes the various technical steps to create a billing history using SAS. Such steps include: 1) identifying a sample of claims and creating a file, 2) writing a SAS program to merge all claims in the universe with those in the sample, 3) selecting just the applicable claims, and 4) creating a report that is informative and easy to read. This chapter is especially helpful for analysts trying to duplicate work done by the Technical Support Staff.

Introduction

This question is asked during many of OEI's Medicare studies. While our inspections tend to focus on a particular type of service, we often have to look at other types of Medicare claims for a sample of beneficiaries to do a complete analysis. We could use a billing history to see if our sample claims look suspicious. For example, if we are looking at a sample of emergency ambulance claims, we may want to check all outpatient claims during the same time period to verify that each beneficiary actually had an emergency room service . If there is no indication that a person ever ended up in an emergency room, we may need to investigate the validity of the sample ambulance claim. We could also use a billing history to compare one group of beneficiaries to another group. For example, if we are looking at a sample of home health services, we may find that beneficiaries that are serviced by company A receive 5 times more visits than beneficiaries that are serviced by company B. We could pull all Medicare claims for these two groups of beneficiaries to determine if company A beneficiaries seem to be sicker than company B beneficiaries.

If the two groups seem to be in similar health, company A may be billing for too many home health visits.

Once we have decided that we need to analyze a billing history, how do we go about doing this? First, we have to identify the sample of claims that we are going to study. For this example, we will limit our focus to claims for HCPCS E0277 (alternating pressure mattress) billed during August, 1996. We have created a file in SAS called "SAMPLE" containing this data.

“SAMPLE” Data File

HCPCS	HICN	ALLOWED	DATE	PLACE
E0277	AAA	100.00	8/01/96	12
E0277	BBB	200.00	8/23/96	32
E0277	DDD	100.00	8/17/96	12

Next we have to create a file (or files) of claims to be included in the billing history that we want to analyze. To do this, we must decide the types of claims we want to include, and what time frame we are going to focus on. For this example, we will look at any support surface claims with HCPCS E0277 (our sample HCPCS), E0180, or E0194. We will also select claims for the sample month (8/96). The result is a SAS filed called "UNIVERSE."

“UNIVERSE” Data File

HCPCS	HICN	ALLOWED	DATE	PLACE
E0277	AAA	100.00	8/01/96	12
E0194	AAA	50.00	8/20/96	12
E0194	AAA	50.00	8/31/96	12
E0180	BBB	25.00	8/02/96	32
E0180	BBB	25.00	8/15/96	32
E0277	BBB	200.00	8/23/96	32
E0180	CCC	25.00	8/03/96	12
E0180	CCC	50.00	8/29/96	12
E0277	DDD	100.00	8/01/96	12
E0277	DDD	100.00	8/17/96	12

Now, we must write a SAS program to pull all claims from the "UNIVERSE" file for the beneficiaries in the "SAMPLE" file.

The first step is to create a new SAS file that contains a list of unique beneficiary HICNs from "SAMPLE" using the following code:

SAS CODE

```
PROC SORT
   DATA=SAMPLE(KEEP=HICN)
   OUT=SORTSAMP;
   BY HICN;
DATA SAMPBENS;
   SET SORTSAMP;
   BY HICN;
   IF FIRST.HICN;
```

END SAS CODE

Suppose we select a sample of three beneficiaries out of the four listed above and put it in a file called "SAMPBENS."

Next, we merge "SAMPBENS" with the "UNIVERSE" file to pull just claims that are for the three sample beneficiaries using the following code:

"SAMPBENS" Data File

HICN
AAA
BBB
DDD

SAS CODE

```
PROC SORT
      DATA=UNIVERSE;
      BY HICN;

DATA BILLHIST;
      MERGE SAMPBENS(IN=SAMP)
            UNIVERSE(IN=UNIV);
      BY HICN;
      IF SAMP AND UNIV THEN OUTPUT;
```

END SAS CODE

The new SAS file "BILLHIST" contains the following data:

"BILLHIST" Data File

HCPCS	HICN	ALLOWED	DATE	PLACE
E0277	AAA	100.00	8/01/96	12
E0194	AAA	50.00	8/20/96	12
E0194	AAA	50.00	8/31/96	12
E0180	BBB	25.00	8/02/96	32
E0180	BBB	25.00	8/15/96	32
E0277	BBB	200.00	8/23/96	32
E0277	DDD	100.00	8/01/96	12
E0277	DDD	100.00	8/17/96	12

Notice that the records for HICN "CCC" were not included since that person was not one of the three sample beneficiaries.

The variables on "BILLHIST" can now be analyzed using various SAS procedures (PROC MEANS, PROC SUMMARY, etc.), but these procedures are beyond the scope of this chapter. Our goal is to demonstrate how to create a billing history and produce a hardcopy for review. The simplest way to produce a hardcopy report of "BILLHIST" is with the SAS PRINT procedure below:

SAS CODE

```
PROC PRINT
DATA=BILLHIST;
BY HICN;
```

END SAS CODE

The output from PROC PRINT would look like this:

```
---------------------- HICN = AAA ----------------------------

  OBS   HCPCS    ALLOWED     DATE     PLACE
   1    E0277    100.00     8/01/96     12
   2    E0194     50.00     8/20/96     12
   3    E0194     50.00     8/31/96     12
---------------------- HICN = BBB ----------------------------
  OBS   HCPCS    ALLOWED     DATE     PLACE
   4    E0180     25.00      8/02/96     32
   5    E0180     25.00      8/15/96     32
   6    E0277    200.00      8/23/96     32
---------------------- HICN = DDD ----------------------------
  OBS   HCPCS    ALLOWED     DATE     PLACE
   7    E0277    100.00       8/01/96     12
   8    E0277    100.00       8/17/96     12
```

In some cases, this type of output is all that we will need. However, there are many occasions for which we may need additional information or a more professional looking report. In older versions of SAS, we had to write a special DATA step program, called DATA _NULL_, to create customized reports. However, the more recent versions of SAS include the REPORT procedure which combines features from the PRINT, MEANS, and TABULATE procedures with features of DATA _NULL_ report writing into one powerful report-writing tool. Because the REPORT procedure will allow you to compute various statistics, group and summarize data, and customize your output, it is a good idea to review SAS documentation (SAS Guide to the REPORT Procedure) for a complete overview of its features. We will walk through one example of how to improve the output created with the PRINT procedure.

But first we must decide what we can do to make our report easier to read and more informative. We could drop the observation number (OBS) if we do not care about that information. We could insert a heading for the report and create more meaningful column headings. We could also move the HICN value into a separate column next to the other variables and get rid of the string of hyphens (-). We also could put a blank line before each HICN and only print the HICN for the first claim so it is easy to tell when one beneficiary ends and another beneficiary begins. It would also be helpful if we could summarize the allowed dollar amount for each beneficiary's claims and display this total after the last claim. Finally, we could format the values of place to provide more meaningful information than "12" or "32".

The REPORT & FORMAT procedures will allow us to do all of these things. However, to change the place values we also have to add a FORMAT procedure to our program. The following SAS code will make the changes that we have indicated above:

SAS CODE

```
  PROC FORMAT;
    VALUE $PLCFMT
     '12'='HOME'
     '32'='NURSING HOME';

  PROC REPORT
    DATA=BILLHIST HEADSKIP SPACING=3;
    TITLE;
    COLUMN ('BILLING HISTORY EXAMPLE' ' ' ' HICN HCPCS
       ALLOWED DATE PLACE);
    DEFINE HICN / ORDER CENTER 'HICN' WIDTH=4
FORMAT=$3.;
    DEFINE HCPCS / DISPLAY CENTER 'HCPCS' WIDTH=5
FORMAT=$5.;
    DEFINE ALLOWED / SUM CENTER 'ALLOWED/AMOUNT'
WIDTH=7
           FORMAT=DOLLAR7.2;
    DEFINE DATE / DISPLAY CENTER 'SERVICE/DATE'
FORMAT=$8.;
    DEFINE PLACE / DISPLAY 'PLACE' WIDTH=12
FORMAT=$PLCFMT.;
    BREAK AFTER HICN / SUMMARIZE SUPPRESS OL SKIP;
```

END SAS CODE

This code performs the following tasks:

1)
```
PROC FORMAT;
VALUE $PLCFMT
'12'='HOME'
'32'='NURSING HOME';
```

This procedure creates a user-defined format called "$PLCFMT" that will be used in the REPORT procedure to print "HOME" every time that PLACE = 12 and "NURSING HOME" every time that PLACE=32.

2)
```
PROC REPORT
DATA=BILLHIST HEADSKIP SPACING=3;
```

Use the REPORT procedure to produce a report. Read data from the SAS file "BILLHIST'. Write a blank line beneath all column headers at the top of each page of the report. Place 3 blank characters between each column of the report.

3)
```
TITLE;
COLUMN ('BILLING HISTORY EXAMPLE' ' ' HICN HCPCS
ALLOWED DATE PLACE);
```

Remove any titles that are in effect (from other SAS procedures, etc.). Add a two-line heading with "BILLING HISTORY EXAMPLE" on the first line and nothing on the second line (insert a blank line). Center this heading above the columns for HICN, HCPCS, ALLOWED, DATE, and PLACE.

4)
```
DEFINE HICN / ORDER CENTER 'HICN' WIDTH=4 FORMAT=$3.;
DEFINE HCPCS / DISPLAY CENTER 'HCPCS' WIDTH=5 FORMAT=$5.;
DEFINE ALLOWED / SUM CENTER 'ALLOWED/AMOUNT' WIDTH=7
        FORMAT=DOLLAR7.2;
DEFINE DATE / DISPLAY CENTER 'SERVICE/DATE' FORMAT=$8.;
DEFINE PLACE / DISPLAY 'PLACE' WIDTH=12 FORMAT=$PLCFMT.;
```

Each variable that will be included in the report must have a DEFINE statement to define its characteristics. This statement indicates how the REPORT procedure should use the variable in the report, formats the variable, and chooses the column header. We are including five variables in this report.

- *HICN, is an ORDER variable (the data is sorted by HICN and will be grouped this way in our report). The HICN data will be under a centered column heading "HICN", the column will be 4 positions wide and HICN is a 3 position character field (the "$" indicates character, without it the variable would be considered numeric).*
- *ALLOWED is a numeric field with the values output in SAS DOLLAR format under the centered two line column heading "ALLOWED AMOUNT." ALLOWED is a SUM variable (in addition to displaying the values for each claim the REPORT procedure will be summarizing their values).*
- *HCPCS, DATE, and PLACE are DISPLAY variables (simply display their data, do not perform any calculations or grouping on these variables). HCPCS and DATE are both character format variables. PLACE has the special format that was defined with the PROC FORMAT.*

5) BREAK AFTER HICN / SUMMARIZE SUPPRESS OL SKIP;

The BREAK statement controls the REPORT procedure's actions when the value of an ORDER variable changes. This line causes the procedure to write a summary line after the last claim for each HICN. This line will only include a sum of the values of ALLOWED for each beneficiary since that is the only variable that was used to compute statistics. SUPPRESS causes the procedure to suppress printing of the HICN value in this summary line. OL writes a line of hyphens (-) above the values in the summary line. SKIP writes a blank line after the summary line.

The output from this code will look like this:

"UNIVERSE" Data File

```
                         BILLING HISTORY
EXAMPLE

                       ALLOWED     SERVICE
 HICN    HCPCS    AMOUNT      DATE        PLACE

 AAA      E0277    $100.00     8/01/96     HOME
           E0194      $50.00        8/20/96  HOME
           E0194      $50.00        8/31/96
                                                         H
                                                         O
                                                         M
                                                         E
              -------
                        $200.00

 BBB     E0180      $25.00        8/02/96
                                                         N
                                                         U
                                                         R
                                                         S
                                                         I
                                                         N
                                                         G
                                                         H
                                                         O
                                                         M
                                                         E
           E0180      $25.00        8/15/96
                                                         N
                                                         U
                                                         R
                                                         S
                                                         I
                                                         N
                                                         G
                                                         H
                                                         O
                                                         M
                                                         E
           E0277    $200.00        8/23/96
                                                         N
                                                         U
                                                         R
                                                         S
                                                         I
```

The billing history is now ready to be analyzed!

CHAPTER 9

EXTRACTING PAYMENTS AND COUNTS BY HCPCS

Analysis Question: How do I want determine the totals for particular Medicare Part B HCPCS?

ABSTRACT

Many studies conducted by OEI involve evaluating Part B services. One of the most important aspects of preinspection is determining how much was paid for various procedure codes of interest. The first step in determining how much was paid involves accessing the data. The second step involves manipulating and printing the data in a manner that provides useful information. This chapter discusses both of these steps by describing four sources of Medicare Part B payment data and how to access and manipulate this data using SAS.

Introduction

Before we start an inspection we often need to know how much Medicare paid for each item or procedure to determine whether a study is worth pursuing. Once we know that we are going to look at a particular area we need to decide which procedure codes (*called HCPCS*) to include in our universe. We have done this for many previous inspections as well as current inspections.

Some examples of inspections where we have had to make this decision are 1) Body Jackets (OEI-04-92-01080), 2) Questionable Medicare Payments for Wound Care Supplies (OEI-03-94-00790), and 3) Support Surfaces, (OEI-02-95-00370). Often there may be a range of HCPCS for a particular service but we may decide to focus on just the big dollar HCPCS such as was done for Body Jackets. If we want to make sure that a specific HCPCS is included in our inspection and there is an uneven distribution with regard to the HCPCS in our population, we have to run totals by HCPCS to determine if we need to stratify before sampling. Totals by HCPCS are also important when analyzing and projecting sample results.

The first question that we need to ask is "where can I find HCPCS level data?". There are four different HCFA sources that we can use to get this data. They are:

1) **BESS** (Part B Extract and Summary System) - Contains summary level Part B data that is updated quarterly.
2) **OIG 1% Sample Files** - Contains 1% of all Medicare Part A and Part B claims. Stored in mainframe SAS files that are updated quarterly by OEI's Technical Support Staff (TSS).
3) **DSAF** (Decision Support Access Facility) - Access to 5% of physician and supplier claims and 100% of institutional claims. Updated quarterly.
4) **MANRLINE** (Method of Accessing Nearline Data) - Access to 100% of physician and supplier claims. This process takes much longer than the other three options (2-4 weeks minimum).

If you only need to know the total services and dollars for several HCPCS, you should use BESS, an easy to use menu-driven system. It contains summary level data only; thus, you are limited in the kinds of information that you can retrieve. It is a useful tool when you are trying to decide if there is enough money being spent in a particular area to justify further analysis. To access BESS, you must logon to the HCFA Data Center and type the following at the TSO "Ready" prompt:

```
EX 'MU00.BESS'
```

You will then be led through a series of menus so that you can define your query. There are options to make a user defined report and options for several "canned" reports. The following steps show a simple example of how to access one of the canned reports.

1) At the primary option menu select #1 (Physician/Supplier Data).
2) The Physician/Supplier data menu will appear. Select #4 (Descriptive Statistics).
3) All of the canned reports will now be listed on the screen. For this example, we will select #8 (type of service by HCPCS).
4) You will be prompted to select a year. All of the totals are summarized by calendar year. Select the year that you want to review.
5) Next you must choose to either browse the output on the screen or to run a batch program that will send the output to a printer or a file. Select the option that you would prefer.
    a. You will then be prompted to enter the HCPCS or range of HCPCS that you are interested in. If you selected "browse" in #5, then the output will appear on the screen. It will appear in a format similar to the one listed below.

```
                                 ALLOWED    ALLOWED   PCT TO
                               AVG ALWD
        HCPCS  DESCRIPTION     SERVICES   CHARGES   TOTAL
   CHARGE

        E0776   PURCHASE DME       7,500         75,000        75.0
10.00
                      RENTAL DME          2,500          25,000
                     25.0       10.00
        TOTAL--ALL PLACES            10,000       100,000       100.0
            10.00
```

b. If you selected "batch processing" in #5, you will be prompted for printer or file information. When the batch program completes, the output (similar to that in #5a) will be sent to the printer or file that you specified.

If BESS data does not meet all of your needs, you will have to retrieve claims data from one of the other three sources. We use the 1% sample files as the basis for most OEI inspections. However, occasionally the 1% sample does not provide enough cases for a procedure or a sampling strata. In those instances, we must use DSAF or MANRLINE to access the larger files. For the rest of this chapter, we will assume that we have extracted our universe of claims from one of these three sources and stored them in a SAS file. This file can be on the mainframe or in PC SAS file.

"SURFACES" Data File

```
HCPCS     HICN   SERVICES
ALLOWED

  E0180     AAA          1
  25.00
  E0180     AAA          2
  50.00
  E0180     CCC          1
  25.00
  E0277     AAA          1
  25.00
  E0277     DDD          2
  75.00
  E1399     BBB         2
100.00
  E1399     BBB         2
100.00
```

For most OEI inspections we keep universe data files on a mainframe because of their large size. We then can download only those cases we select for use on the PC. To keep things simple for this example, we will assume that we have created a SAS file called "SURFACES."

There are several ways that you can use SAS to summarize this data. We will cover two possible ways in this chapter: 1) SUMMARY Procedure and 2) DATA Step.

The SUMMARY Procedure computes statistics on numeric variables in a SAS file and outputs the results to a new SAS file. We can use this procedure to summarize our claims data by total services and total allowed dollars by HCPCS. To do this, we can use the following SAS code:

SAS CODE

```
PROC SUMMARY
DATA=SURFACES;
CLASS HCPCS;
VAR SERVICES ALLOWED;
OUTPUT OUT=TOTALS
        SUM(SERVICES)=TOTSRVC
        SUM(ALLOWED)=TOTALLW;
```

END SAS CODE

This code creates a SAS file called "TOTALS" which contains four records (see Table).

"TOTALS" Data File

```
  HCPCS    TOTSRVC    TOTALLW
_FREQ_ _TYPE_

               11          400.00
        7              0
  E0180         4          100.00
```

There is one record for each HCPCS, and the new variables TOTSRVC and TOTALLW contain a sum of the values of the variables SERVICES and ALLOWED on the "surfaces" file. The other two variables (_FREQ_ and _TYPE_) are created by SAS every time that you run a PROC SUMMARY. _FREQ_ is a count of the number of records that went into each subgroup. For example, there were three records on "SURFACES" that had HCPCS "E0180" so the _FREQ_ value for E0180 is 3. The values of _TYPE_ indicate which subgroup (defined by the combinations of the class variables) produced the summary statistics in that record. A _TYPE_ value of 0 indicates that the record is a total summary of all of the records on the input file. In this case, there are seven records on "surfaces" (_FREQ_=7) which had 11 total services and 400.00 total allowed dollars. A _TYPE_ value of 1 indicates that the statistics are for the first class variable. Since we only have one class variable (HCPCS), this indicates that the values are for each HCPCS. An example of a PROC

SUMMARY with two class variables is given later in this chapter.

When performing certain statistical operations on a file, such as computing a mean, do not include the summary record because it will distort the calculation. If you decide that you do not want your output file (totals) to have a summary record (_TYPE_=0), you can use the NWAY option on the PROC SUMMARY statement. NWAY specifies that statistics should be output for only the records with the highest _TYPE_ value (highest level of interaction among CLASS variables). Also, you can use a DROP statement if you decide that you do not want to keep the _TYPE_ or _FREQ_ variables on your output file. These two changes to the SAS code are shown below.

SAS CODE

```
PROC SUMMARY
DATA=SURFACES NWAY;
CLASS HCPCS;
VAR SERVICES ALLOWED;
                           OUTPUT
                           OUT=TOTALS1(DROP=_TYPE_
                           _FREQ_)
                           SUM(SERVICES)=TOTSRVC
                           SUM(ALLOWED)=TOTALLW;
```

END SAS CODE

"TOTALS1" Data File

HCPCS	TOTSRVC	TOTALLW
E0180	4	100.00
E0277	3	100.00
E1399	4	200.00

Now, the output SAS file (TOTALS1) only contains three records (HCPCS,TOTSRVC, and TOTALLW).

By default, PROC SUMMARY does not produce printed output like most of the other SAS procedures. However, you can use the PROC PRINT to print the output file that PROC SUMMARY creates. Adding these lines to the end of our SAS code will produce printed output.

SAS CODE

```
PROC PRINT
DATA=TOTALS1;
TITLE 'PROC SUMMARY OUTPUT FOR
SURFACES';
```

END SAS CODE

Looking back at the data on our original file (SURFACES), we see that we were able to summarize the SERVICES and ALLOWED variables by HCPCS using PROC SUMMARY. However, the HICN data was not used. Using the HICN, we can get a count of the number of beneficiaries that had claims for each HCPCS. Yet, PROC SUMMARY will not allow us to compute statistics on this variable because it is not numeric. However, we can make HICN a CLASS variable as shown below.

SAS CODE

```
 PROC SUMMARY
 DATA=SURFACES;
CLASS HCPCS HICN;
VAR SERVICES ALLOWED;
OUTPUT OUT=TOTALS2
SUM(SERVICES)=TOTSRVC
SUM(ALLOWED)=TOTALLW;
```

END SAS CODE

The result is a SAS file called "TOTALS2" which contains 13 records.

"TOTALS2" Data File

```
  HCPCS    HICN    TOTSRVC    TOTALLW
_FREQ_    _TYPE_

                 11              400.00    7
                                                  0
                  AAA        4           100.00
                                     3         1
           BBB        4          200.00
                                                  2



                                                  1
             CCC        1            25.00
                                                  1
```

As you can see, SAS provides a summary of each interaction among CLASS variables. We have the overall total (_TYPE_=0), the records for only the four HICNs totaled, the records for each of the three HCPCS totaled, and the combination of each HCPCS with each HICN. In this case, we do not care about the overall totals (_TYPE_=0), the totals for each HICN (_TYPE_=1), or the totals for each HCPCS (_TYPE_=2). We are only concerned with the highest level of interaction between CLASS variables (NWAY), the HCPCS*HICN totals (_TYPE_=3). If we use the same adjustments as those we made to our previous program, NWAY option and the DROP statement, our SAS code will look like this:

SAS CODE

```
PROC SUMMARY
DATA=SURFACES NWAY;
CLASS HCPCS HICN;
VAR SERVICES ALLOWED;
OUTPUT OUT=TOTALS3(DROP=_TYPE_
_FREQ_)
SUM(SERVICES)=TOTSRVC
SUM(ALLOWED)=TOTALLW;
```

END SAS CODE

This code creates an output file (TOTALS3) containing five records.

We now have a file with one record for each HCPCS HICN combination. Remember that when you run a PROC SUMMARY. SAS creates a variable called _FREQ_ that contains a count of the number of records that went into each subgroup. Therefore, if we input this file into another PROC SUMMARY with one CLASS variable (HCPCS), the _FREQ_ variable that is created will represent the total number of beneficiaries for each HCPCS. This is reflected in the code below. We will also add a RENAME statement to make the _FREQ_ variable name more meaningful.

"TOTALS3" Data File

```
HCPCS     HICN   TOTSRVC
          TOTALLW

E0180     AAA        3           7
                                 5
                                 .
                                 0
                                 0
E0180     CCC        1           2
                                 5
                                 .
                                 0
                                 0
E0277     AAA        1           2
                                 5
                                 .
                                 0
                                 0
E0277     DDD        2           7
                                 5
                                 .
                                 0
                                 0
E1399     BBB        4
                              200.00
```

SAS CODE

```
PROC SUMMARY
DATA=SURFACES NWAY;
CLASS HCPCS HICN;
VAR SERVICES ALLOWED;
OUTPUT OUT=TOTALS(DROP=_TYPE_ _FREQ_)
 SUM(SERVICES)=TOTSRVC
SUM(ALLOWED)=TOTALLW;

 PROC SUMMARY
 DATA=TOTALS NWAY;
```

```
CLASS HCPCS;
VAR TOTSRVC TOTALLW;
OUTPUT OUT=NEWTOTS(DROP=_TYPE_
RENAME= (_FREQ_=TOTBENES))
SUM=;
```

END SAS CODE

This code creates a new SAS file (NEWTOTS) which has one record for each HCPCS. By typing "SUM=;" we have asked SAS to summarize the variables in the VAR statement and keep the same names on the output file. Therefore "newtots" has a variable called TOTSRVC and another called TOTALLW. The output is shown below.

This program will work as long as your input file is not too large. SAS will only allow up to 32,767 combinations to be output from a PROC SUMMARY. Therefore, if you have a file with claims for more than 32,767 benes, SAS will not allow you to include the HICN as a CLASS variable. For files larger than 32,767, it is better to write your own summary routine using a SAS DATA step. To do this, first sort the data with the SAS SORT procedure. The following code will do the same thing as the two PROC SUMMARY routines above.

"NEWTOTS" Data File

```
  HCPCS    TOTSRVC    TOTALLW
TOTBENES

   E0180        4          100.00
                          2
   E0277        3          100.00
                          2
   E1399        4          200.00
                                      1
```

SAS CODE

```
PROC SORT
DATA=SURFACES
OUT=SRT;
BY HCPCS HICN;

DATA NEWTOTS(KEEP=HCPCS TOTBENES
TOTSRVC
      TOTALLW);
```

```
SET SRT;
BY HCPCS HICN;
RETAIN TOTBENES TOTSRVC TOTALLW;
IF FIRST.HCPCS THEN DO;
        TOTBENES=0;
        TOTSRVC=0;
        TOTALLW=0;
END;
IF FIRST.HICN THEN TOTBENES+1;
TOTSRVC+SERVICES;
TOTALLW+ALLOWED;
IF LAST.HCPCS THEN OUTPUT;
```

END SAS CODE

This program sorts the data in "SURFACES" by HCPCS and HICN and places it in a new SAS file "srt". For this example, we could have sorted the data and rewritten it to "surfaces." However, you are usually only going to use this type of program for a large input file. Therefore, it is a good practice to write the dataset out to a new file to avoid problems with space allocation. Once the file is sorted, the code in the DATA step performs the following functions:

1) DATA NEWTOTS(KEEP=HCPCS TOTBENES TOTSRVC TOTALLW);

> *This line creates a new SAS file called "newtots" and tells SAS that this file should only contain the 4 variables that are listed in the KEEP statement.*

2) SET SRT;
BY HCPCS HICN;

> *Use the file "srt" as input to this routine. The file is sorted by HCPCS and HICN.*

3) RETAIN TOTBENES TOTSRVC TOTALLW;

> *Causes the values of TOTBENES, TOTSRVC, and TOTALLW to retain their values from one input record to the next. They are only changed when a specific SAS statement changes them (when they are assigned a new value). The SAS default does not retain a value from one record to the next; so, without this statement the values of these variables would be reset every time that a record was read from "srt".*

4) IF FIRST.HCPCS THEN DO;

```
TOTBENES=0;
TOTSRVC=0;
TOTALLW=0;   END;
```

When the first occurrence of each HCPCS value is read from "srt," initialize the variables TOTBENES, TOTSRVC, and TOTALLW to 0. Otherwise, ignore these statements.

5) IF FIRST.HICN THEN TOTBENES+1;

When the first occurrence of each HCPCS HICN combination is read from "srt" add 1 to the variable TOTBENES. Otherwise, ignore this statement.

6) TOTSRVC+SERVICES;

Every time that a record is read from "srt," add the value of SERVICES to the value currently retained in TOTSRVC and store this new total value in TOTSRVC.

7) TOTALLW+ALLOWED;

Every time that a record is read from "srt," add the value of ALLOWED to the value currently retained in TOTALLW and store this new total value in TOTALLW.

8) IF LAST.HCPCS THEN OUTPUT;

When the last occurrence of each HCPCS has been read from "srt" and all the other statements in the DATA step have been processed, output the values of HCPCS, TOTBENES, TOTSRVC, and TOTALLW to file "newtots". Otherwise, ignore this statement. This is the only time that anything will be written to "newtots." This allows us to input seven records from "srt" and output three records to “newtots."

The output file (NEWTOTS) contains three records and is identical to the output that was created by using two PROC SUMMARY routines.

There are also other procedures in SAS such as PROC TABULATE and PROC MEANS that can give you summary information for a HCPCS. These procedures are beyond the scope of this chapter.

"NEWTOTS" Data File

```
  HCPCS    TOTSRVC    TOTALLW
TOTBENES

  E0180       4             100.00
                                 2
  E0277       3             100.00
                                 2
  E1399       4             200.00
                                 1
```

APPENDIX A

STATISTICS: An Overview

The purpose of our research, as with most any type of research, is to assess relationships between and among a set of variables. Research can be classified as one of three types: *experimental, quasi-experimental, or observational.*

Experiments are the most controlled type of study and maximize the investigator's ability to isolate the observed effect of the dependant variables from the distorting effects of the independent variables. When observational units are assigned randomly to treatment and control groups, the study is considered to be an experiment.

Quasi-experiment are often more feasible and less expensive than experiments but offer less control over the study. When observational units are assigned to treatment groups *without* randomization, the study is considered to be quasi-experimental.

Observational experiments are the easiest studies to implement but offer the least potential for drawing definitive conclusions. When all observations are obtained without either randomization or comparison groups, then the study is considered to be observational.

In The Office of Inspector General (OIG), we typically encounter observational studies primarily because we are bound by limited time and budget constraints. We are interested in determining what relationships play a significant role in various outcome variables. We use several different types of statistical techniques to accomplish this.

CLASSIFICATION OF VARIABLES

Variables can be classified a number of ways. These classifications are useful in determining the method of data analysis we will use. I'm going to describe three methods of variable classifications: gaped/not gaped, descriptive orientation, and level of measurement.

DESCRIPTIVE ORIENTATION

In the this section we are going to talk about whether a variable is to describe or be described by other variables. If a variable under investigation is to be described in terms of other variables, we call it a response, or *dependant* variable. If we are using a variable in conjunction with other variables to describe a given response variable, we call it a predictor, or *independent*, variable. Other variables that may affect the relationships between dependant and independent variables but have no intrinsic value in a particular study are referred to as control or nuisance variables. Also, in some contexts, these variables are referred to as covariates or confounders. It is important to note that a variable considered as dependent for evaluating one study objective may be considered independent for evaluating another study objective.

LEVELS OF MEASUREMENT

The third classification scheme deals with the precision of measurement of the variable. There are three of these levels. They are: nominal, ordinal, and interval.

NOMINAL

The weakest level of measurement is nominal. At this level the values assumed by a variable usually indicate different categories. The variable 'sex' or 'gender' is nominal because we assign the numbers 1 and 0 to denote female and male, respectively, to distinguish the two categories.

ORDINAL

A somewhat higher level of measurement allows us not only to group into separate categories but also to order the categories. This level is called ordinal. Social class is an ordinal variable since an ordering can be made among the different classes. An ordinal scale possess all the properties of a nominal scale plus ordinality.

INTERVAL

A variable that can not only give ordering but also give a meaningful measure of the difference between categories is an interval variable. To be interval, a variable must have a well-accepted physical unit of measurement. Height, weight, blood pressure are all examples of interval variables while subjective measures like personality type, social class, and stress levels do not. Rates of occurrence are also

An interval variable that has a scale with a true zero is occasionally designated as a ratio, or ratio-scale, variable. An example of a ratio scale variable is the height of a person. Temperature measured in degrees Celsius, an interval scale while measurement of temperature in kelvin is referred to as absolute zero, and so is a ratio variable. An example of a ratio variable common in the health industry is the concentration of a substance in the blood (i.e., cholesterol). Rates of occurrence are also ratio level data, for example, HMO disenrollment rates or hospital admissions.

BASIC STATISTICS: AN INTRODUCTION

Statistics is one of the four mathematical sciences and concerns the methods and procedures for collecting, classifying, summarizing, and analyzing data. The primary goal of most statistical analysis is to make statistical inferences, that is, draw valid conclusions about a population based on information contained in a sample of that population. Therefore, the applications of statistics can be divided into two broad areas; descriptive and inferential statistics.

- Descriptive statistics - utilizes numerical and graphical methods to look for patterns,
  to summarize, and to present information about a set of data.

- Inferential statistics - utilizes sample data to make estimates, decisions, predictions, and other generalizations about a larger set of data sometimes referred to as a population of elements.

DESCRIPTIVE STATISTICS

In this area of statistics there are several tools available to assist us in describing a data set. Graphs are example of descriptive statistics. Some graphs that are frequently used are Stem and Leaf plots, Histograms, Box and Whisker plots, and Circle Graphs, or Pie Charts. Because most computer programs construct these charts automatically, it is beyond the scope of this introduction to describe the steps to construct each one. It is important that the reader understand that the pictorial methods of describing a data set can be very useful in determining the distribution of your data set (i.e., normal distribution, skewed distribution) and also help in determining what method of data analysis is useful.

Another commonly used tool in descriptive statistics are numerical measures of Central Tendency. When we talk about a data set, we are talking about a sample or a population. Since inference is our ultimate goal, we want to use numerical descriptive values from the sample to make inferences about the corresponding population. There are lots of numerical methods available to describe data sets and most measure one of two data characteristics:

1. The central tendency of the set of measurements, or the tendency of the data to *center* or *cluster* around a particular value.
2. The variability of the set of measurements, or the *spread* of the data.

NUMERICAL MEASURES OF CENTRAL TENDENCY

In this section I'm going to concentrate on measures of central tendency. The most popular and best understood measure of central tendency for a quantitative data set is the *arithmetic mean* or simply the *mean*.

- The mean of a set of quantitative data is equal to the sum of values divided by the number of values contained in the data set.

In everyday terms, the mean is the average value of the data set.

The median is another important measure of central tendency. The median is of most value in describing large data sets and in working with nonparametric data sets. That is data sets that do not satisfy the assumptions that the *t*- and *F*- tests are based on.

- The median is the middle number when the measurements in a data set are arranged in ascending (or descending) order. If *n* is odd, the median is the middle number. If *n* is even, the median is the mean of the middle two numbers.

In some situations the median is a better measure of central tendency than the mean because the median is less sensitive to extremely large or small measurements. The median is useful to compare to the mean to get a rough idea of the shape of your data set. The following is a quick Rule of Thumb when comparing the median and the mean.

- If the median is less than the mean, the data set is skewed to the right.

- If the median is greater than the mean, the data set is skewed to the left.

- If the median is equal to the mean, the data set is symmetrical.

A third measure of central tendency is the mode. The mode is often used with large data sets to locate the region where most of the data is concentrated. The mode is used primary when describing very large data sets.

- The mode is the measurement that occurs most frequently in the data set.

NUMERICAL MEASURES OF VARIABILITY

Measures of central tendency only give us a small part of the information that we need when want to describe a quantitative data set. The description is incomplete without a measure of variability, or spread, of the data set. Two data sets can have the same mean, and yet still be very different because one may be more spread out than the other. Perhaps the simplest measure of the variability of a quantitative data set is the range.

- The range of a data set is equal to the largest measurement minus the smallest measurement.

The range is easy to compute and easy to understand but it is pretty insensitive to actual variation in a data set when the data set is large. This is easy to understand because two data sets can have the same range but be extremely different with respect to the actual variation.

The two measures of variability most often considered are the sample variance and the sample standard deviation.

- The sample variance is equal to the sum of the squared distances of each value of a variable from the mean divided by (n - 1). In symbols, we use s^2 to represent the sample variance.

The drawback to using s^2 is that it is in squared units of the variable x. To have a measure of dispersion that is expressed in the same units as x, we simply take the square root of s^2 and call it the sample standard deviation.

- The sample standard deviation, s, is defined as the positive square root of the sample variance, s^2.

When we compare the variability of two samples selected from a population, the sample with the larger standard deviation is the more variable of the two. Unfortunately, none of us have the time to pull two samples each time to make variability comparisons. There are two rules that we can use to act as an aid to the interpretation of a standard deviation. The first applies to any set of data and is derived from a theorem proven by Russian Mathematician, Chebyshev.

CHEBYSHEV'S RULE

Chebyshev's Rule applies to any sample of measurements, regardless of the shape of the frequency distribution.

1. It is possible that very few measurements will fall within 1 standard of the mean.

2. At least 3/4 of the measurements will fall within 2 standard deviations of the mean.

3. At least 8/9 of the measurements will fall within 3 standard deviations of the mean.

4. Generally, at least $1 - 1/k^2$ of the measurements will fall within k standard deviations of the mean for any number, k, greater than 1.

The second rule which applies only to mound-shaped distributions of data, is based on empirical evidence that has accumulated over time. It is appropriately called The Empirical Rule.

EMPIRICAL RULE

The Empirical Rule is a rule of thumb that applies to samples with frequency distributions that are mound-shaped.

1. Approximately 68% of the measurements will fall within 1 standard deviation of the mean.

2. Approximately 95% of the measurements will fall within 2 standard deviations of the mean.

3. Essentially all of the elements will fall within 3 standard deviations of the mean.

I've given you several ways that we can use when we are interested in describing a quantitative set of data. It is important that the reader understand that the area of descriptive statistics is broader and deals with several different types of distributions not presented here (for example, binomial, logistic, and Chi-Square distributions). That information, although very important, is beyond the scope of this short introduction and the reader is encouraged to review random variables and their distributions for clarification of those issues.

INFERENTIAL STATISTICS

As we discussed in the beginning, the primary goal of most statistical analysis is to draw valid conclusions about a population based on information gathered from a sample of that population. There are two general categories of statistical inference, they are; estimation and hypothesis testing. Each one of these categories has a different purpose. Estimation is concerned with estimating the specific value of an unknown population parameter. Hypothesis testing is concerned with making a decision about a hypothesized value of an unknown population parameter. Let's talk about these two in a little more detail.

ESTIMATION

In estimation we want to estimate an unknown parameter using a random variable. This point estimator is in the form of a formula, or rule (i.e., mean, or standard deviation).

The usual procedure is to select a random sample from the population of interest, calculate the point estimate (i.e., the mean), and then associate a measure of variability with it, which usually takes the form of a confidence interval.

> A confidence interval (CI) consists of two boundary points between which we have a certain specified level of confidence that the actual population parameter lies.

For example, a 95% confidence interval for a parameter would consist of upper and lower limits determined, so that in repeated sets of samples of the same size, 95% of all the intervals would be expected to contain the true population parameter.

HYPOTHESIS TESTING

We use our the estimation process to develop what we think is a likely set of values for the parameters of interest. Next, we use the hypothesis testing process to test whether our estimated value for the parameter is different enough from some hypothesized value (usually referred to as the null hypothesis) to conclude that the hypothesized value is not likely to be true.

The general procedure used in testing a statistical null hypothesis is basically the same regardless of the parameter being considered. The procedure consists of the following seven steps:

1. Check the assumptions of the properties of the underlying variables being tested to insure that the testing procedure is appropriate.

2. State the null hypothesis *Ho*, and the alternative hypothesis, *Ha*.

3. Specify the level of significance, the alpha level.

4. Specify the test statistic to be used, *t-*, *F-*, *Chi-square*, and its distribution under *Ho*.

5. Form the decision rule for rejecting and not rejecting the null hypothesis.

6. Compute the value of the test statistic from the observed data.

7. Draw your conclusions concerning the rejection or nonrejection of the null hypothesis.

There are some other points of interest when you are involved in hypothesis testing is what is called a P-value, and Type I and Type II errors. I'll briefly discuss them for you now. The P-value is a value that we can compute that quantifies *exactly how unlikely the observed results would be if Ho were true.* Another way to describe the P-value is as follows:

- The P-value gives the probability of obtaining a value of the test statistic at least as unfavorable to Ho as the observed value.

It is important to note that you can use P-values can be used to draw conclusions about a test. If you decide to use P-values as the basis for accepting or rejecting the null hypothesis, the following guidelines are recommended:

1. If *P* is small (less than .05), reject Ho.

2. If *P* is large (greater than .5), do not reject Ho.

3. If $.05 < P < .5$, the significance is borderline; that is we reject Ho for an alpha equal to .1 but not for an alpha equal to .01.

Note: If we actually do specify alpha a priori, we reject Ho when *P* is less than alpha.

There are actually two types of error that can be made when performing a statistical test: A Type II error occurs if we fail to reject Ho when Ho is actually false. We denote the probability of a Type II error by *beta* and call (1 - *beta*) the power of the test.

For a fixed sample size, alpha and beta for a given test are inversely related, which means that lowering one has the effect of raising the other. It is important to note that in general, the power of any statistical test can be raised by increasing the sample size.

In this section, we have discussed the overall concepts of research and the different ways of classifying our variables. We have talked about the difference between descriptive statistics and inferential statistics and how to apply those techniques to a data set of interest. These are only guidelines to help give you a deeper understanding of why we are doing what we're doing in the quantitative analysis arena.

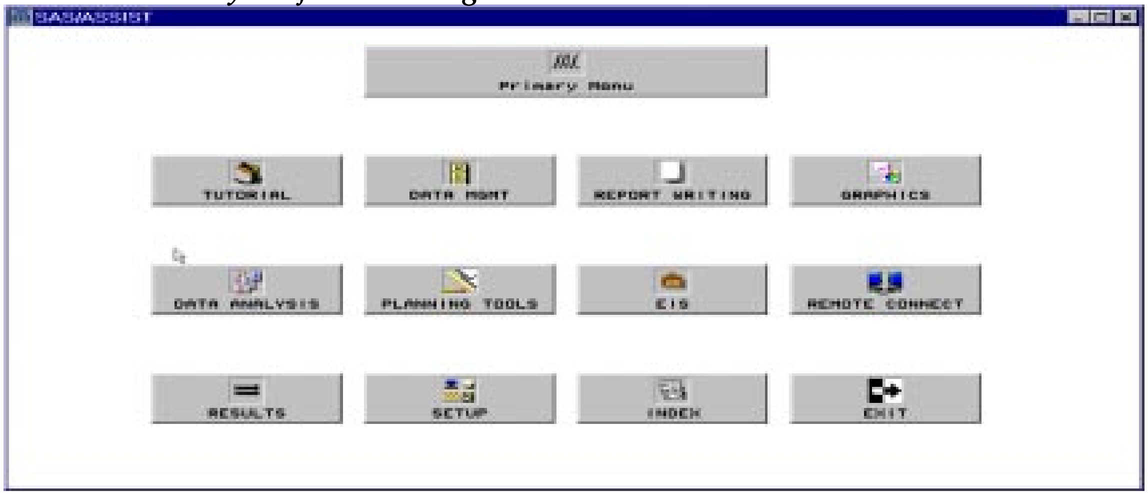
APPENDIX B

USING SAS WINDOWS ASSIST

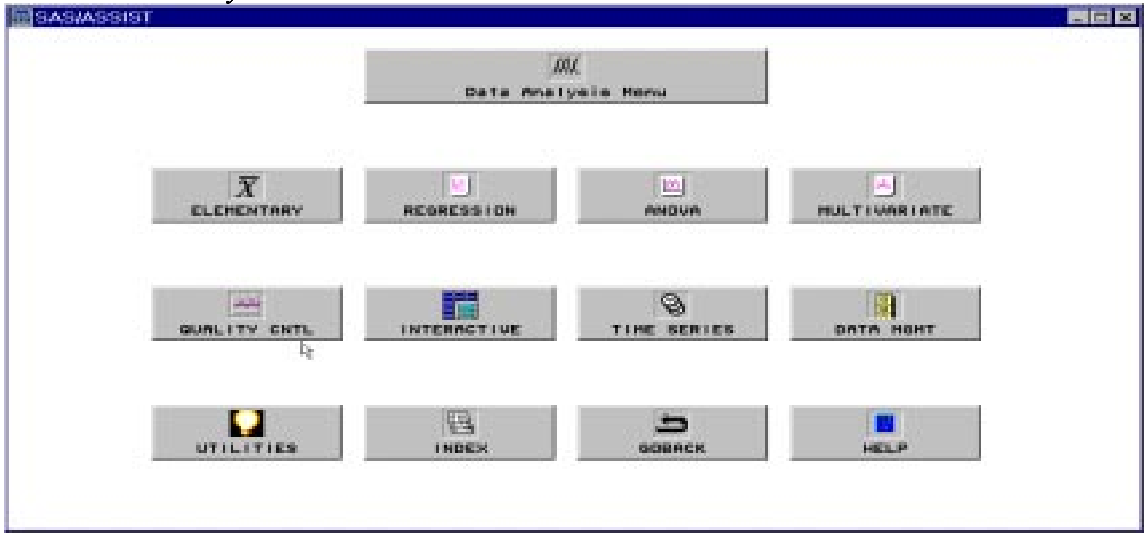
An Example: How to Calculate Confidence Intervals
With SAS Assist

Creating confidence intervals is even easier with SAS Assist software - simply select from the menu options. The following screens show what you would see when using SAS assist.

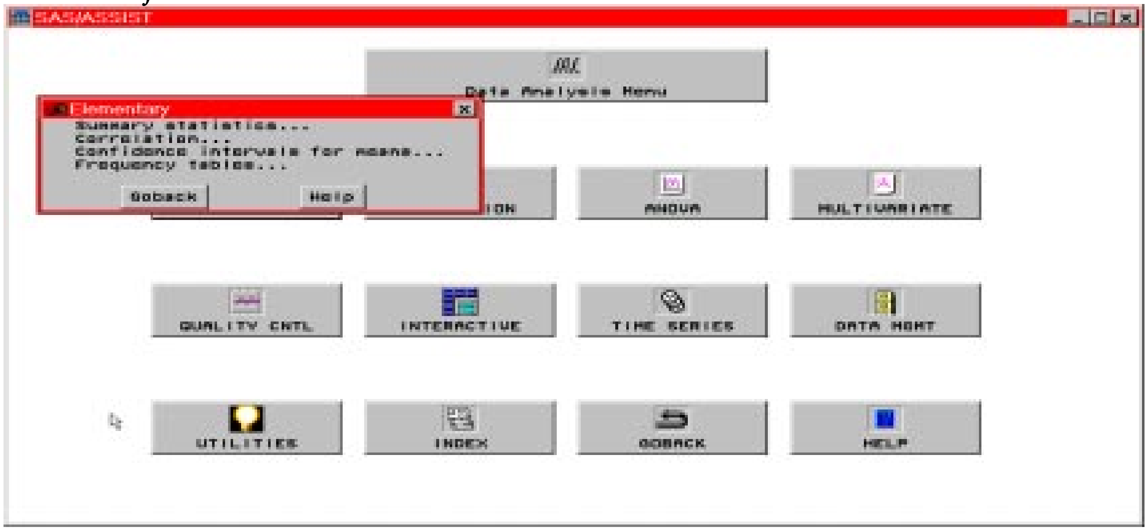
Select Data Analysis after executing the Assist module.



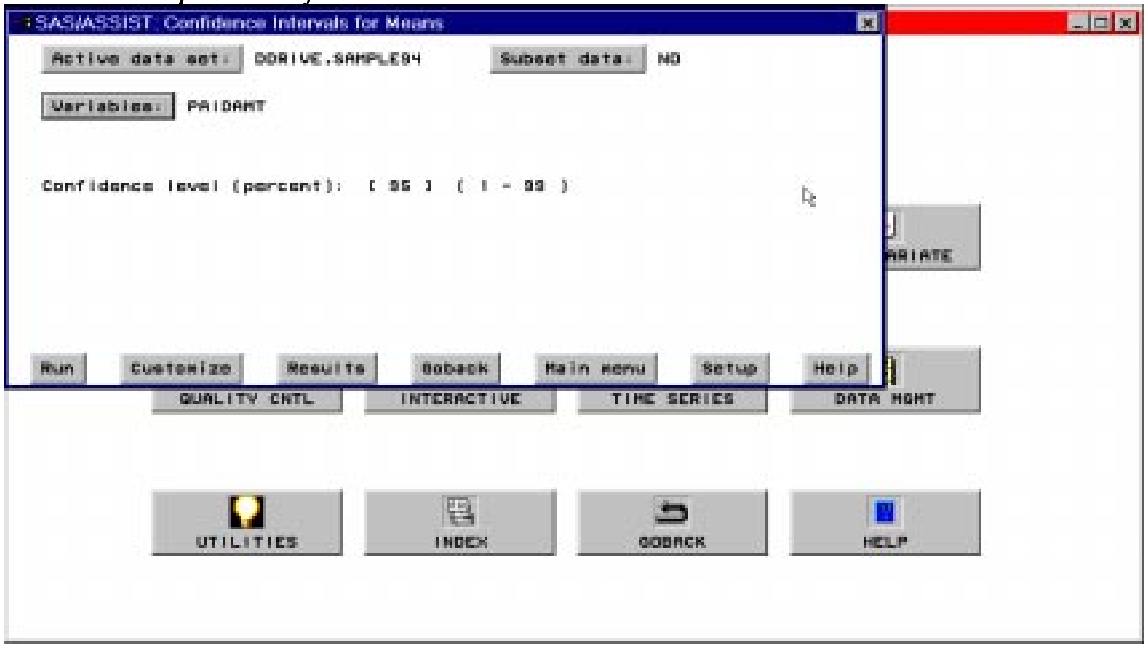
Select Elementary



Select Confidence Intervals



Fill in the requested information



Active data set: DDRIVE.SAMPLE94

Subset data: NO

Variables: PAIDAMT

Confidence level (percent): [95] (1 - 99)

OUTPUT

```
    VARIABLE    MEAN      DF     LEVEL     LOWERCL     UPPERCL

     PAIDAMT   25.4722  89986     95        25.2367     25.7078
```

PROGRAM CREATED BY SAS ASSIST

The following is the program that SAS assist created. Note this program does not utilize the new Proc Means options.

```
  OPTIONS LINESIZE=80 PAGESIZE=63 DATE NUMBER PAGENO=1;
  TITLE;
  FOOTNOTE;
  PROC MEANS NOPRINT DATA=DDRIVE.SAMPLE94;
   VAR PAIDAMT;
   OUTPUT OUT =SASAST1
       N   =N1
       MEAN=MEAN1
       STD =STD1;
  RUN;
  %LET _ASERR = &SYSERR;
  DATA SASAST2;
   SET SASAST1;
   NAMELIST = "PAIDAMT";
   ARRAY NUM   { 1 } N1;
   ARRAY AVG   { 1 } MEAN1;
   ARRAY STDV  { 1 } STD1;
   HALFALF = 1 - ( (1 - .95) / 2 );

   DO I = 1 TO 1;
    LEVEL = 95;
    DF = NUM{ I } - 1;
    IF ( DF < 0 ) THEN DF = . ;
    MEAN = AVG{ I };
    EST = TINV( HALFALF , DF ) * STDV{ I } / SQRT( NUM{ I } );
    LOWERCL = AVG{ I } - EST;
    UPPERCL = AVG{ I } + EST;
    VARIABLE = SCAN( NAMELIST , I , ' ' );
    OUTPUT;
    END;
RUN;
PROC PRINT DATA=SASAST2 NOOBS;
 VAR VARIABLE MEAN DF LEVEL LOWERCL UPPERCL;
RUN;
PROC DATASETS;
 DELETE SASAST1 SASAST2;
RUN;
QUIT;
```

GLOSSARY

STATISTICAL TERMS FOR REFERENCE

Adjusted R² - An R² (R-squared) adjusted to give a truer (smaller) estimate of how much the independent variable in a regression analysis explain the dependent variable. The adjustment is made by taking into a - count the number of independent variables. The adjusted R² is a measure of strength of association. Also called "epsilon-squared."

Alpha Error - An error made by rejecting a true null hypothesis (such as claiming that a relationship exists when it does not). Also called Type I Error.

Alpha Level - (a) The chance a researcher is willing to take of committing an alpha error or Type I Error, that is, of rejecting a null hypothesis that is true. (b) The probability that a Type I Error (wrongly rejecting the null hypothesis) has been committed. The smaller the alpha level, the more significant the finding because the smaller the chance that the finding is due to chance alone. Thus an alpha level of .01 is a more difficult criterion to satisfy than a level of .05. Also called level of (statistical) significance.

Alternative Hypothesis - In hypothesis testing, any hypothesis alternative to the one being tested, usually the opposite of the null hypothesis. Also called the research hypothesis. Rejecting the null hypothesis shows that the alternative (or research) hypothesis *may* be true. Symbolized: H_1 or H_a.

Analysis of Variance (ANOVA) - A test of the statistical significance of the differences among the mean scores of two or more groups on one or more variables or factors. It is an extension of the *t* test, which can only handle two groups, to a larger number of groups. More specifically, it is used to for assessing the statistical significance of the relationship between categorical independent variables and a continuous dependent variable. The procedure in ANOVA involves computing a ratio (*F* ratio) of the variance within the groups (error variance) to the variance between the groups (explained variance).

Association, Statistical - (a) A relationship between two or more variables that can be described statistically. (b) Any of several statistical techniques (such as correlations and regression analysis) that can be used to describe the degree to which differences in one variable are accompanied by (associated with) corresponding differences in another variable.

Association, Test of - Another term for test statistic. Not to be confused with a *measure* of association, which indicates the size of the relation between two variables. By contrast, a *test* of association gives the probability that an association of a given size could have occurred by chance, that is, whether it is statistically significant.

Bell-Shaped Curve - A symmetrical curve, usually plotting a continuous frequency distribution, such as a normal distribution, which looks like a cross section of a bell. The Student's *t* distribution is also bell-shaped, although it is rarely referred to that way.

Best Linear Unbiased Estimator - A regression line computed using the least squares criterion when none of the assumptions are violated. Abbreviated: BLUE.

Beta Coefficient - A regression coefficient for a sample expressed in standard deviation units (i.e., *z*-scores). Specifically, the beta coefficient indicates the difference in a dependent a dependent variable associated with an increase (or decrease) of one standard deviation in an independent variable -- when controlling for the effects of other independent variables. Also called standardized regression coefficient and beta weight. Note: A regression coefficient expressed in nonstandardized units is usually symbolized by *b*. Usage is confusing because beta is also used to symbolize the population parameter of *b*.

Beta Error - An error made by accepting or retaining a false null hypothesis -- more precisely, by *failing to reject* a false null hypothesis. This might involve, for example, claiming that a relationship does not exist when it in fact does. Also called Type II Error. Compare alpha error.

Bias - (a) Anything that produces systematic error in a research finding. More formally, bias is the difference between the expected value of a sample statistic and the population parameter the statistic estimates. (b) Also, the effects of any factor that the researcher did not expect to have an influence on the dependent variable. Compare random error.

Binomial Distribution - A probability distribution for a dichotomous or two-value variable (binomial = "two-names"), such as success/failure, profit/loss, or in/out. Also called "Bernoulli distribution."

Categorical Variable - A variable that distinguishes among subjects by putting them into a limited number of categories, indicating type or kind, as sex does by categorizing people into male or female. Also called "discrete" or "nominal" variable. Compare continuous variable.

Central Limit Theorem - A statistical proposition to the effect that, the larger a sample size, the more closely the sampling distribution of the mean will approach a normal distribution. This is true even if the population from which the sample is drawn is not normally distributed. A sample size of 30 or more will usually result in a sampling distribution of the mean that is very close to a normal distribution. The central limit theorem explains why sampling error is smaller with a large sample than it is with a small sample.

Central Tendency, Measure of - Any of several statistical summaries of data designed to find a single number that best represents several numbers. Examples include the mean, the mode, and the median.

Chi-Square Distribution - A family of theoretical probability distributions, each of which has a different degree of freedom. The chi-square test is based on it.

Chi-Square Test - A test statistic, that is, one used to assess the statistical significance of a finding.

Coefficient - (a) A number used as a measure of a property or characteristic. (b) In an equation, a number by which a variable is multiplied.

Cohort - A group of individuals having a statistical factor (usually age) in common. Compare a social category. For example, all persons born in 1996 form a cohort.

Confidence Interval - A range of values of a sample statistic that is likely (at a given level of probability, called a confidence level) to contain a population parameter. The interval that will include the population parameter a certain percentage (confidence level) of the time. The wider the confidence interval, the higher the confidence level.

Confidence Level - A desired percentage of the scores (usually 95%) that will fall within a certain range of confidence limits. It is calculated by subtracting the alpha level from 1 and multiplying the result times 100; for example, 100 x (1 - .05) = 95%.

Confidence Limits - The upper and lower values of a confidence interval, that is, the values defining the range of a confidence interval.

Correlation Coefficient - A number showing the degree to which two variables are related. Correlation coefficients range from -1.0 to +1.0.

Covariance - A measure of the joint or (co-) variance of two or more variables.

Covariation - (a) A state that exists when two things -- such as the price and the sales of a commodity -- vary together.

Critical Region - The area in a sampling distribution representing values that are "critical" to a particular study. They are critical because when a sample statistic falls in that region, the researcher can reject the null hypothesis.

Deduction - (a) A conclusion that follows logically from known (or assumed) principles, that is, that uses deductive methods. (b) The process of reasoning that moves from general principles to conclusions about particular instances.

Degrees of Freedom - Usually abbreviated "df." The number of values free to vary when computing a statistic. This number is necessary to interpret a chi-square statistic, an F ratio, and a t score.

Denominator - Another term for the divisor; in division, the part of the fraction that is below the line.

Dependent Variable - (a) The presumed effect in a study; so called because it "depends" on another variable. (b) The variable whose values are predicted by the independent variable, whether or not caused by it.

Descriptive Statistics - Procedures for summarizing, organizing, graphing, and, in general, describing quantitative information. Often contrasted with inferential statistics, which is used to make inferences about a population based on information about a sample drawn from that population.

Determination, Coefficient of - A statistic that indicates how much of the variance in one variable is determined or explained by one or more other variables; more strictly, how much the variance in one is associated with variance in the others. It is calculated by squaring the correlation coefficient. Thus it is abbreviated r^2 in bivariate analyzes and R^2 in multivariate analyzes. Also called "index of determination." For example, one might find a statement like the following in a research report: "Education level attained explains 22% of adult occupational status (r^2 =, *.22).*"

Dichotomous Variable - A categorical variable that can place subjects into only two groups, such as male/female, alive/dead, or pass/fail.

Difference of Proportions - A method for comparing proportions for dichotomous variables. One proportion is subtracted from the other. The result ranges from -1.0 to +1.0, with zero indicating that the two variables have identical conditional probabilities on a dependent variable.

Discrete Variable - Commonly, another term for categorical (or nominal) variable. Compare continuous variable.

Dispersion, Measure of - A statistic showing the amount of variation or spread in the scores for, or values of, a variable. When the dispersion is large, the scores or values are widely scattered; when it is small, they are tightly clustered. The two most commonly used measures of dispersion are the variance and the standard deviation.

Dummy Variable - A dichotomous variable, usually coded 1 to indicate the presence of an attribute and 0 to indicate its absence. Example: 1 = female; 0 = not female. This coding facilitates the use of interval-level statistical techniques, which would be hard to interpret if the variable were coded otherwise, such as female = 2, male = 1.

Efficiency - (a) In experimental design, said of a procedure that uses fewer resources for the same results or that gets more results using the same resources. (b) In statistics, a property of an estimate of a population parameter; the better the estimate, the greater the efficiency. Efficiency is a measure of the variance of an estimate's sampling distribution; the smaller the variance, the better the estimator.

Endogenous Variable - A variable that is an inherent part of the system being studied and that is determined from within the system. In other words, a variable that is caused by other variables in a causal system. Generally contrasted with exogenous variable.

Error - The difference between an observed score and a predicted or estimated score. Symbolized as *e* or *E*.

Error Sum of Squares - In analysis of variance or regression, the within-group sum of squares, that is, the part not explainable by the treatment effects or regression model.

Error Term - The part of an equation indicating what is unexplained by the independent variables.

Estimation - Using a sample statistic to determine the probable value of a population parameter.

Exogenous Variable - A variable entering from and determined from outside the system being studied. A causal system says nothing about its exogenous variables.

Expected Value - (a) The mean value of a variable in repeated samplings or trials. (b) The mean of the sampling distribution of a statistic.

External Validity - The extent to which the findings of a study of a sample may be generalized to a population.

Extrapolation - Inferring values by projecting trends beyond known evidence.

F - (a) Uppercase *F*, the statistic that is computed when conducting an analysis of variance.

***F* Test** - A test of the results of a statistical analysis, perhaps most closely associated with, but by no means limited to, analysis of variance (ANOVA). The *F* test yields an *F* ratio or *F* statistic. This is a ratio of the variance between groups (explained variance) to the variance within groups (unexplained variance). To tell whether the *F* ratio is statistically significant, you have to consult an *F* distribution table.

Goodness-of-Fit Test - The chi-square test applied to a single categorical variable to see if the distribution among categories matches (fits) a theoretical expectation. The bigger the chi-square statistic, the poorer the fit; the smaller, the better.

H_0 - The symbol for the null hypothesis.

H_a - A symbol for the alternative or research hypothesis.

Homogeneous - Generally, the same or similar.

Induction - Using statistical methods to form generalizations by finding similarities among a large number of cases. The generalizations derived in this way are probabilistic. For example, if 90% of the members of the U.S. Congress were lawyers, the chances of any individual member of the Congress being a lawyer would be 9 out of 10.

Kurtosis - The shape (degree of peakedness) of a curve that is a graph representation of a (unimodal) frequency distribution. Kurtosis usually indicates the extent to which a distribution departs from the bell-shaped or normal curve by being either pointier (leptokurtosis) or flatter (platykurtosis).

Lagged Dependent Variable - Said of a dependent variable whose value at a particular time is to some degree dependent on its value at a previous time.

Level of Significance - More fully, the level of statistical significance. The probability that a result would be produced by chance (sampling error, random error) alone. The level of significance indicates the risk or probability of committing an error (Type I Error in hypothesis testing). The level of significance is stated as a probability, often abbreviated *p*, followed by a number, for example, $p \leq .05$ or $p > .01$. The smaller the number, the smaller the chance of Type I Error and the more statistically significant the finding.

Logistic Regression Analysis - A kind of regression analysis used when the dependent variable is dichotomous and scored 0, 1. It is generally used for predicting whether something will happen or not, such as graduation, business failure, heart disease -- anything that can be expressed as Event/Nonevent. Independent variables may be categorical or continuous in logistic regression analysis. It is based on transforming data by taking their natural logarithms so as to reduce nonlinearity. Rather than using OLS methods, logistic regression estimates parameters using maximum likelihood estimation.

Mean - The average. To get the mean, you add up the values for each case and divide the total by the number of cases.

Median - The middle score in a set of ordered scores. When the number of scores is even, there is no single middle score; in that case, the median is found by taking an average of the two middle scores.

Mode - The most common (most frequent) score in a set of scores.

Multicollinearity - In multiple regression analysis, multicollinearity exists when two or more independent variables are highly correlated; this makes it difficult if not impossible to determine their separate effects on the dependent variable.

Multiple Linear Regression - A method of regression analysis that uses more than one predictor variable (or independent variable) to predict a single dependent variable. The coefficient for any particular predictor variable is an estimate of the effect of that variable while holding constant the effects of the other independent variables.

Multivariate Analysis - Any of several methods for examining multiple variables at the same time.

N - Number. Usage varies; among the most common meanings of the uppercase *N* are (a) number of subjects or cases in a particular study, (b) number of individuals in a population, © number of variables in a study.

n - Number. Usage varies; among the most common meanings of the lowercase *n* are (a)n number in a sample, as opposed to in a population, (b) number of cases in a subgroup. For example, consider the following from a research report: "We interviewed a random sample of college graduates ($N = 520$) to get their opinions on several issues; males were 45% ($n = 234$) of the sample." This means that a total of 520 graduates were interviewed; 234 of them were in the male subgroup.

Nested Variables - Said of variables located inside other variables -- such as city, state, and national murder rates.

Nominal Scale (or Level of Measurement) - A scale of measurement in which numbers stand for names but have no order or value. For example, coding female = 1 and male = 2 would be a nominal scale; females do not come first; two females do not add up to a male, and so on. The numbers are merely labels.

Nominal Variable - Another term for a categorical (or a discrete or a qualitative) variable. See nominal scale.

Normal Distribution - A purely theoretical continuous probability distribution in which the horizontal axis represents all possible values of a variable and the vertical axis represents the probability of those values occurring. The scores on the variable (often expressed as *z*-scores) are clustered around the mean in a symmetrical, unimodal pattern known as the bell-shaped curve or normal curve. In a normal distribution, the mean, median, and mode are all the same. There are many different normal distributions, one for every possible combination of mean and standard deviation. Also sometimes called the "Gaussian distribution." Because the sampling distribution of a statistic tends to be a normal distribution, the normal distribution is widely used in statistical inference. For small samples, the Student's *t* distribution (which is also "bell-shaped" but not "normal") is used.

Null Hypothesis - (H_0) The hypothesis that two or more variables are *not* related or that two or more statistics (e.g., means for two different groups) are not the same. In accumulating evidence that the null hypothesis is *false*, the researcher indirectly demonstrates that the variables *are* related or that the statistics are different. The null hypothesis is the core idea in hypothesis testing.

Numerator - In a fraction, the number above the line; the number into which the denominator is divided.

Odds Ratio - A ratio of one odds to another. The odds ratio is a measure of association, but, unlike other measures of association, "1.0" means that there is no relationship between the variables. The size of any relationship is measured by the difference (in either direction) from 1.0. An odds ratio less than 1.0 indicates an inverse or negative relation; an odds ratio greater than 1.0 indicates a direct or positive relation. Also called "cross-product ratio."

One-Tailed Test of Significance - A hypothesis test stated so that the chances of making a Type I (or alpha) Error are located entirely in one tail of a probability distribution.

Ordinal Scale (or Level of Measurement) - A scale of measurement that *ranks* subjects (puts them in an order) on some variable. The differences between the ranks need not be equal (as they are in an interval scale). Team standings or scores on an attitude scale (highly concerned, very concerned, concerned, and so on) are examples.

Ordinary Least Squares (OLS) - A statistical method of determining a regression equation. That is, the equation that best represents the relationship among the variables, given the criteria of minimizing the sum of squares of the residual (error term) between the predicted value and the observed value.

Outlier - A subject or other unit of analysis that has extreme values on a variable. Outliers are important because they can distort the interpretation of data or make misleading a statistic that summarizes values (such as a mean).

Oversampling - A procedure of stratified sampling in which the researcher selects a disproportionately large number of subjects from a particular group (stratum). Most often, researchers oversample in a stratum that has a large variance or in a stratum that would yield too few subjects if a simple random sample were used.

P **- Probability value, or *p* value.** Usually found in an expression such as *p .05*. This expression means: "The probability (*p*) that this result could have been produced by chance (or random error) is less than (<) five percent (.05)." Thus, the smaller the number, the greater the likelihood that the result expressed was not merely due to chance. For example, *p* < *.001* means that the odds are a thousand to one (one tenth of 1%) against the result being a fluke. What is being reported (.05, .002, and so on) is an alpha level or significance level. The *p* value is the actual probability associated with an obtained statistical result; this is then compared with the alpha level to see whether that value is (statistically) significant.

Point Estimate - An estimate made by computing a statistic that describes a sample; this is then used to estimate a population parameter.

Poisson Distribution - A probability distribution used when the number *(N)* of cases is very large and the probability *(p)* is very small.

Population - A group of persons (or institutions, events, or other subjects of study) that one wishes to describe or about which one wishes to generalize. To generalize about a population, one often studies a sample that is meant to be representative of the population. Also called "universe."

Population Parameter - A characteristic of a population described by a statistic, such as a mean or a correlation. Population parameters are usually symbolized by Greek letters; the Roman (English) alphabet is often used for sample statistics. Consistency is not perfect here, however, as Greek letters are sometimes used for both statistics and parameters.

Power of a Test - Broadly, the ability of a technique, such as a statistical test, to detect relationships. Specifically, the probability of rejecting a null hypothesis when it is false -- and therefore should be rejected. The power of a test is calculated by subtracting the probability of a Type II Error from 1.0. The maximum total power a test can have is 1.0; the minimum is zero. Also called "statistical power."

Probability Level - The *p* value below which the null hypothesis is rejected; this value or level is the chance of making a Type I (or alpha) Error.

Probit Analysis - A technique used in regression analysis when the dependent variable is a dummy (or dichotomous) variable. It assumes a cumulative normal distribution in contrast to the logistic distribution assumed in Logit analysis.

Proportional Stratified Random Sample - A stratified random sample in which the proportion of subjects in each category (stratum) is the same as in the population.
p Value - Short for probability value.

Qualitative - (a) When referring to variables, "qualitative" is another term for categorical or nominal.

Quantitative - Said of variables or research that can be handled numerically.

Quartiles - Divisions of the total cases or observations in a study into four groups of equal size (quarters).

Random Sampling - Selecting a group of subjects (a sample) for study from a larger group (population) so that each individual (or other unit of analysis) is chosen entirely by chance.

Random Variation - Differences in a variable that are due to chance rather than to one of the other variables being studied.

Rank Order Scale - Another term for an ordinal scale, that is, one that gives the relative position of a score in a series of scores.

Regression - Any of several statistical techniques concerned with predicting some variables by knowing others.

Regression Coefficient - A number indicating the values of a dependent variable associated with the values of an independent variable or variables. A regression coefficient is part of a regression equation.

Reliability - The consistency or stability of a measure or test from one use to the next.

Robust - Said of a statistic that remains useful even when one (or more) of its assumptions is violated. For example, the *F* ratio is generally robust to violations of the assumption that treatment groups have equal variances.

Sample - A group of subjects selected from a larger group in the hope that studying this smaller group (the sample) will reveal important things about the larger group (the population).

Sampling Error - The inaccuracies in inferences about a population that come about because researchers have taken a sample rather than studied the entire population. In other words, sampling error is the difference between a population parameter and a sample statistic used to estimate that parameter. Sampling error is one kind of random error.

Sampling Fraction - The size of a sample as a percentage of the population from which it was drawn; the ratio of sample size to population size.

Sampling Frame - A list or other record of the population from which the sampling units are drawn.

SAS - Statistical Analysis System. A widely used statistical package for data analysis in the social and behavioral sciences.

Significance - The degree to which a research finding is meaningful or important.

Significance Level - The probability of making a Type I Error. The lower the probability, the higher the statistical significance.

Significance Testing - Using statistical tests (such as chi-square, *t* test, or *F* test) to determine how likely it is that observed characteristics of samples have occurred by chance alone in the populations from which the samples were selected. If the observed characteristics in the samples are unlikely to be due to chance alone, the characteristics are deemed statistically significant.

Simple Regression - A form of regression analysis in which the values of a dependent variable are attributed to (are a function of) a single independent variable.

Skewed Distribution - A distribution of scores or measures that, when plotted on a graph, produce a nonsymmetrical curve. In a unimodal skewed frequency distribution, the mode, mean, and median are different. When the skewness of a group of values is zero, their distribution is symmetrical.

Specification Error - A mistake committed when deciding upon (specifying) the causal model in a regression analysis. The three most common such errors are (1) leaving an important variable out of the causal model (2) including an irrelevant variable and (3) using the wrong functional form of a variable (i.e. using X when X^2 is needed).

Standard Deviation - A statistic that shows the spread or dispersion of scores in a distribution of scores; in other words, a measure of dispersion. The more widely the scores are spread out, the larger the standard deviation. The standard deviation is calculated by taking the square root of the variance. It is symbolized as SD or *s* or as a lowercase sigma.

Standard Error - Often short for standard error of the mean or standard error of estimate. The smaller the standard error, the better the sample statistic is as an estimate of the population parameter -- at least under most conditions. The standard error is a measure of sampling error; it refers to error in our estimates due to random fluctuations in our samples. The standard error is the standard deviation of the sampling distribution of a statistic.

Standard Error of the Mean - A statistic indicating how greatly the mean score of a single sample is likely to differ from the mean score of a population. It is the standard deviation of a sampling distribution of the means.

Standardized Regression Coefficient - A statistic that provides a way to compare the relative importance of different variables in a multiple regression analysis. It measures the impact of a one standard deviation change in a regression coefficient.

Standard Normal Deviate - Another term for standard score or *z*-score.

Standard Score - A measure of relative standing in a group arrived at by transforming raw scores in a way that allows one to compare raw scores from different distributions. The common standard score is the *z*-score.

Statistic - A number that describes such characteristic of a variable or of a group of data -- such as a mean or a correlation coefficient.

Statistical Inference - Using probability and information about a sample to draw conclusions ("inferences") about a population or about how likely it is that a result could have been obtained by chance.

Statistical Power - A gauge of the sensitivity of a statistical test, that is, its ability to detect effects of a specific size, given the particular variances and sample sizes of the study.

Statistical Significance - Said of a value or measure of a variable when it is ("significantly") larger or smaller than would be expected by chance alone.

- It is important to remember that statistical significance does not necessarily imply substantive or practical significance. A large sample size very often leads to results that are statistically significant, even when they might be otherwise quite inconsequential.

Stratified Random Sampling - Random or probability samples drawn from particular categories (or "strata") of the population being studied.

Stratum - A subgroup of population, based on a selected criteria (i.e. male, female).

Student's *t* Distributions - A family of theoretical probability distributions used in hypothesis testing. The *t* distribution is used for interpreting data gathered on small samples when the population variance is unknown.

Sum of Squared Errors - In a regression analysis, what you are trying to minimize when you use the ordinary least-squares criterion. This is calculated by summing the squares of the difference between the observed dependent variable values and those predicted by the estimated regression equation.

Sum of Squares - The result of adding together the squares of deviation scores.

***T*-Score** - (uppercase T) A way of expressing deviation from a mean.

***t* Statistic** - The number that is tested in a *t* test, that is, the number that is compared with the critical region. It is used for small sample inference.

***t* Test** - A test of the statistical significance of the results of a comparison between two group means, such as the average score on a manual dexterity test of those who have and have not been given a caffeine drink.

Two-Tailed Test of Significance - A statistical test in which the critical region (region of rejection of the null hypothesis) is divided into two areas at the tails of the sampling distribution. This test is used when we are concerned whether the mean of one group is higher or lower than the other.

Type I Error - An error made by wrongly rejecting a true null hypothesis.

Type II Error - An error made by wrongly accepting (or retaining or failing to reject) a false null hypothesis.

Unbiased Estimate - A sample statistic that is free from any systematic bias leading it to over- or underestimate the corresponding population parameter.

Universe - Another term for population.

Validity - A term to describe a measurement instrument or test that measures what it is supposed to measure; the extent to which a measure is free of systematic error.

Variability - The spread or dispersion of scores in a group of scores; the tendency of each score to be unlike the others.

Variance - A measure of the spread of scores in a distribution of scores, that is, a measure of dispersion. It is calculated by summing the square of each observed value minus the mean and dividing this sum by N-1 (sample data).

Weighted Average (or Mean) - A procedure for combining the means of two or more groups of different sizes; it takes the sizes of the groups into account when computing the overall or grand mean.

Weighted Data - (a) Any information given different weights in calculations, as when the final examination counts twice as much as (is weighted double) the midterm. (b) Data whose values have been adjusted to reflect differences in the number of population units that each case represents.

***z*-Score** - (lowercase z) - The most commonly used standard score. In *z*-score notation, the mean is 0 and a standard deviation is 1. Thus a *z*-score of 1.25 is one and one-quarter standard deviations above the mean. It is used when the underlying population has a normal distribution of the measured variable.

REFERENCES

SELECTED STATISTICAL PUBLICATIONS

BASIC TEXTS

Blalock, Hubert M., Jr. 1979. Social Statistics: Revised Second Edition. New York, New York: McGraw-Hill Book Company.

Babbie, Earl. 1995. The Practice of Social Research: Seventh Edition. Belmont, California: Wadsworth Publishing Company.

Cody, Ronald P. And Jeffrey K. Smith. 1987. Applied Statistics and the SAS Programming Language. New York, New York: Elsevier Science Publishing Co.,Inc.

Freund, John E. And Frank J. Williams. 1991. Dictionary/Outline of Basic Statistics. Mineola, New York: Dover Publications.

Gonick, Larry and Woollcott Smith. 1993. The Cartoon Guide to Statistics. New York, New York: Harper Collins Publishers, Inc.

Levy, Paul S. And Stanley Lemeshow, Ph.D. 1980. Sampling for Health Professionals. Belmont, California: Wadsworth Publishing Company,

Mendenhall, William. 1994. Introduction to Probability and Statistics: Nine Edition. Belmont, California: Duxbury Press, a division of Wadsworth Publishing Company.

Scholotzhauer, Sandra D. And Ramon C. Littell. 1987. SAS System for Elementary Statistical Analysis. Cary, North Carolina: SAS Institute, Inc.

Vogt, Paul W. 1993. Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences. Newbury Park, California: SAGE Publishing, Inc.

ADVANCED READING

Aldrich, John and Forrest D. Nelson. 1984. Linear Probability, Logit, and Probit Models. Newbury, California: Sage Publications, Inc.

Cochran, William G. 1977. Sampling Techniques: Third Edition. New York, New York: John Wiley and Sons, Inc. -New York, New York.

Draper N. R. and H. Smith. 1966. Applied Regression Analysis. New York, New York: John Wiley and Sons, Inc.

Hatcher, Larry, Ph.D. and Edward J. Stepanski, Ph.D., 1994. A Step-by-Step Approach to Using the SAS System for Univariate and Multivariate Statistics. Cary, NC: SAS Institute Inc.

Kish, Leslie. 1965. Survey Sampling. New York, New York: John Wiley and Sons, Inc.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. 1953. Sample Survey Methods and Theory. New York, New York. John Wiley and Sons, Inc.

Schroeder, Larry D., David L. Sjoquist, and Paula E. Stephan. 1986. Understanding Regression Analysis: An Introductory Guide. Beverly Hills, California: Sage University Papers.

Snedecor, George W. and William G. Cochran. 1989. Statistical Methods. Ames, Iowa: Iowa State University Press.