Oracle Utilities Load Analysis

Sampling Package User's Guide

Release 1.11.0.3 for Windows

E18233-05

May 2013

ORACLE

Oracle Utilities Load Analysis Sampling Package User's Guide, Release 1.11.0.3 for Windows

E18233-05

Contents

What's New

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Chapter 8

Chapter 9

Chapter 10

Chapter 11

Appendix A

Appendix B

What's New

New Features in the Oracle Utilities Load Analysis Sampling Program

This chapter outlines the new features of the 1.11.0.0 release of the Oracle Utilities Load Analysis that are documented in this guide.

New Features for Release 1.11.0.0

Feature	Description	For more information, refer to...
Record Definition Utility Removed	The record definition utility has been removed from the application.	**How To Create the Record Definition File** on page 3-8.
Population Data File is Delimited	The population data file is now in a delimited file format.	**Guidelines for Population Data File (PDF) Creation** on page 3-10
Buffer Parameter Removed	The Buffer parameter in the sampling parameters file is not supported by this version of the application. The SAMPLING_BUFFERSIZE setting in the CSLSTAR.GLB file replaces the Buffer parameter.	**CSLSTAR.GLB (Run Time)** on page 1-6 of the *Oracle Utilities Load Analysis Configuration Guide*

Chapter 1

Defining Objectives and Sample Design Criteria

Introduction

Welcome to the Oracle Utilities Load Analysis Sampling Package, a set of software programs that aid utility analysts and statisticians in the creation of statistically-reliable samples for any type of customer study, including load research and mail surveys. The package provides the tools necessary for *all* phases of the sampling process — including sample design, selection, and validation. For the design phase, three alternative techniques are supported — simple random, single dimensional stratified, and multidimensional stratified.

The package uses any customer data file or prior survey data as input, and produces sample designs to meet desired confidence levels. It streamlines the sampling process by performing calculations automatically, and provides a great deal of flexibility for matching sample designs to utility requirements.

What Kind of Samples Can You Create With the Sampling Package?

Because the package can incorporate any combination of up to seven *usage and demographic* variables in a single sample design, it can be used to create samples for a broad range of load and market research applications.

For example, you can design samples to collect load data and estimate demand characteristics for user-specified customer groups. You can identify the groups by any combination of variables such as usage (kW or kWh for a particular time or period, for instance), rate class, SIC code, geographic territory, building type, income, family size, and/or any other customer characteristic known for the entire population. Such load studies are useful for cost allocation, rate setting, forecasting, demand side management, conservation, and other utility programs.

You can also design samples of customer groups for mail surveys, to measure appliance saturation (the percentage of customers having a specific appliance) and to gain insight into customer attitudes and preferences. These surveys are useful for identifying demographic trends that will affect energy consumption, and for gauging customer receptivity to conservation and marketing programs.

What You Should Know Before Getting Started

This manual describes, step-by-step, how to use the features and functions of the Sampling Package to create samples tailored to your specific requirements. This document is not intended to teach you the basics of using your computer and operating system. If you need help with this, contact your facility's IT department.

The manual also assumes a basic understanding of statistical sampling, and some familiarity with the terms and concepts of load research. If you are new to load research, you might find it helpful to read the *Oracle Utilities Load Analysis Load Data Management User's Guide* and *Oracle Utilities Load Analysis Load Data Analysis User's Guide.*

How This Manual is Organized

This manual is divided into six major parts according to the "phases" of working with the package, and the appendices.

Chapter 1: **Defining Objectives and Sample Design Criteria** through **Chapter 3**: **Creating the Population Data File and Record Definition File** detail the planning, decision-making, and setup tasks you must perform before you can begin to actually design a sample with the Sampling Package software.

Chapter 2: **Selecting a Sample Design Technique and Establishing a Plan** explains and compares the three sampling techniques supported by the Sampling Package, to help you decide which is most appropriate for any project at hand.

Chapter 4: **Analyzing the Population Frequency Distribution** and **Chapter 5**: **Stratifying the Population and Determining Sample Size for a Single Dimension** explain the steps for designing a single dimensional stratified or simple random sample.

Chapter 6: **Assigning the Population to Cells and Calculating Population Statistics** and **Chapter 8**: **Determining Sample Size for a Multidimensional Sample Design** describe the steps required to design a multidimensional stratified sample.

Chapter 9: **Selecting the Sample for a Single Dimensional Design** and **Chapter 10**: **Selecting the Sample for a Multidimensional Design** tell you how to select your sample, and **Chapter 11**: **Validating the Sample** explains how to validate it.

Appendix A: **The User Language** provides an in-depth explanation of the User Language — a powerful command language you will use to create files and reports. **Appendix B**: **Sampling Equations** documents the calculations used in the Sampling Package.

This manual describes the inputs and outputs of each program. In this package, the outputs of one program are used as inputs for the next. For a description of this data flow, see **Chapter 2**: **Selecting a Sample Design Technique and Establishing a Plan**.

Conventions Used in This Manual

The formats for creating input files are illustrated in boxes throughout the manual. Within these boxes, the following conventions apply:

- Keywords, which you will enter in the file, appear in the manual as a combination of upper- and lower-case letters. Typically, you need enter only the first three letters, which appear in upper-case.
- Parameters you will enter appear in italic.
- Braces { } are used to indicate a choice of parameters, from which you must choose one.
- Brackets [ ] are used to indicate optional parameters which may or may not be specified.
- Vertical bars | separate mutually exclusive choices.
- Default parameter values are indicated by underlines.

Near the beginning of most chapters, you will find a "summary box" that briefly outlines each of the steps required to complete a specific task. Within these boxes, the following naming conventions are used:

ELEMENT	CONVENTION	EXAMPLE	NOTE
Input/Output Files	TGB_ _ _	TGB22A = Population Analysis Control File	File reference recognized by programs

How to Get Help...

Occasionally, you may encounter an error message or other problem. If the message is not self-explanatory, you may be able to decipher it using the *Oracle Utilities Load Analysis Configuration Guide.*

If you cannot find a solution in the *Oracle Utilities Load Analysis Configuration Guide*, you can contact Oracle Support personnel at http://metalink.oracle.com. My Oracle Support offers you secure, real-time access to Oracle experts on the complete Oracle Utilities Load Analysis system. It also provides ground breaking personalized & proactive support capabilities that help reduce unplanned down time and improve system stability. Leverage the Internet for immediate access to 24/7 support and get the critical and timely information you need for running your business.

Before contacting, please prepare the following:

- All outputs associated with the job.
- All applicable Environment and/or Control Files.
- File structure of your Population Data File (PDF) and the PDF joboutput.
- Anything else that you think might aid in diagnosing the problem.

File Placement

All files created in sampling are stored in the Common/Data/Sampling directory. Control Files created within each Sampling Program will be stored in their respective program folder. For example, a control file created for B110 Record Definition Program will be stored in Common/Data/Sampling/B110 folder.

Files that are shared across sampling programs are stored at the root (Common/Data/Sampling) folder. The following files are stored in the Common/Data/Sampling folder:

- Record Definition File (.**DEF**)
- Population Data File (.**PDF**)
- Scored Population Data File (.**PDS**)
- Sampling Parameter File (.**SPF**)
- Frequency Distribution File (.**FDF**)
- Relative Accuracy File (.**RAF**)
- Population Statistics File (.**PSF**)
- Sample Statistics File (.**SSF**)
- Report Control File (.**RCF**)

Sampling Projects

When a Sampling Project is created (see Chapter 3), the new sampling project folder will be created under Common/Data/Sampling. For example, if a new Sampling Project, "2008 Sampling", is created, a new Common/Data/Sampling/2008 Sampling folder will be created. All subsequent sampling files will be stored under this folder.

Sampling Return Codes

1 is returned from the Sort Routine and indicates a problem with the data to be sorted.

This condition could occur when no records are selected by stratification step in B410/B420. Check the stratification control file for proper selection criteria.

11 means a Sampling build sort routine error. Please call Oracle Utilities Corporation.

12 means required keywords/values were not found in the Sampling Parameter File or are invalid. See The Sampling Parameter File (SPF) in chapter 3.

99 means an unrecoverable error occurred. Please review the report for details.

Working with the Sampling Package Software

Careful planning in the early stages of a sampling project can help ensure maximum success. This chapter discusses key questions that must be asked and answered before you actually begin working with the Oracle Utilities Load Analysis Sampling Package software. These questions are summarized below and discussed in detail throughout the remainder of this chapter.

KEY QUESTIONS FOR ESTABLISHING SAMPLE DESIGN CRITERIA

- What are the objectives of the sample?
- What is the target population?
- What information is already available about the population?
- What are the characteristics of interest?
- What is the desired accuracy?
- What measure of variance will be used to determine sample size?
- Is data from a prior study available?
- What are the restrictions (budget, time constraints, etc.)?

Once you have carefully answered these questions, you will have a better understanding of your population and will be able to select the most appropriate sampling technique and possibly several alternative test scenarios. Sampling techniques and test scenarios are discussed further in **Chapter 2**: **Selecting a Sample Design Technique and Establishing a Plan**.

Key Questions

- **What are the objectives of the sample?**

In other words, how will the results be used? The answer to this question will influence your entire sampling program.

Traditionally, load research has been conducted primarily to provide information for rate setting and cost of service studies. The requirements for such studies are often fairly straightforward (and in some cases mandated by law) — typically rate class demand at the hour of system peak, with 90% confidence and 10% accuracy.

However — as utilities have come to recognize the value of load and demographic data as a means of better understanding and serving their customers in a competitive environment, and as sampling tools have become more sophisticated, analysts are being asked to design more creative samples that will meet the needs of a variety of departments throughout the utility. For example:

> Demand side management groups typically want usage profiles for specific appliances, customers, or groups of customers. Their aim might be to identify opportunities for load shifting or reduction, or to evaluate the success of load management programs.
>
> Marketing groups would also be interested in load profiles for groups of customers as a means of targeting customers for potential energy sales or demand side management program participation. They might also request a sample for other types of customer studies, such as a mail survey. Since the cost to send out a letter is vastly different from the cost to install a load research recorder, you will have a great deal more latitude in designing the sample.

The Oracle Utilities Load Analysis Sampling Package gives you the flexibility to design samples to meet any of these needs. As these examples point out, it is very important that you work closely with the department or departments requesting the information to clearly define the objectives. The nature of these objectives will dictate the answers to the remaining questions.

- **What is the target population?**

You must clearly identify the specific group for which information is desired (i.e., the "target population"). The target population might be defined using one or more of the following breakdowns:

Residential, commercial, and industrial classifications.

Rate classes within a classification.

Specific customer groups within a class (defined by SIC Code, end-use applications, annual energy use, geographic location, or other criteria).

Most often, a sample is designed for one customer class at a time. However, using the multidimensional design approach, it is possible to collect data for different customer sub-groups within a single sample.

Whatever definition you use, selection of the population must be based on unambiguous and consistent rules. And, whatever characteristics you use to define your target population must be identified for every member of the population.

> **Note:** Throughout this manual, we refer to the sampling unit as a "customer". Your definition of a sampling unit will be more specific. For example, it might be a customer billing account or a customer premise. Other possible sampling units might include transformers, whole office buildings, and even individual appliances.

- **What data is already available?**

At the outset of the project, it is helpful to take a general inventory of the information already available to you. A great wealth of data exists in utilities (billing files, prior studies, appliance surveys, and more) and additional data is available from outside sources, such as other utilities and marketing research firms that sell demographic profiles. All of these data sources can be very useful to you for a variety of purposes.

For example, the Customer Information System will typically be the source from which the sample population is drawn; but you may consider virtually any reliable source. You can even combine data from different sources.

Any data that exists for all customers (or whatever the sampling unit may be) can be considered as a candidate for a design variable, for instance, SIC Code, geographic location, usage data, etc.

Prior load data is especially important because it can provide estimates for the target variable that may be unavailable for the current population. Incorporating survey data can improve accuracy and efficiency (this is explained further under the question about variance).

- **What are the characteristics of interest?**

You must also carefully identify the characteristics of interest: the data to be measured for the sample population and estimated from the target population. In the case of load research, this is most often demand at a particular hour or demand over a specified day or day-type. Typical characteristics of interest for load research applications are:[1]

Rate class demand at the hour of class monthly, seasonal, or annual peak.

Rate class demand at the hour of system monthly, seasonal, or annual peak.

Hourly rate class demands for the day of the class monthly, seasonal, or annual peak.

Hourly rate class demands for the day of the system monthly, seasonal, or annual peak.

Hourly rate class demands for the average day, weekday, and weekend day defined by month or season.

---

1. From Load Research Manual, Association of Edison Illuminating Companies, 1990.

Class diversity or coincidence factors and load factors.

Total rate class energy usage by day, month, season, or year.

Hourly demand for end-use appliances.

Using the multidimensional sampling approach, it is possible to design a sample that incorporates multiple demand characteristics. For example, you might design a sample that yields the desired accuracy for both winter and summer peaks.

- **What is the desired accuracy?**

The accuracy requirements are determined by the objectives of the study. For example, samples for rate setting require more accuracy than those for simple marketing surveys. PURPA specified a design accuracy of $\pm$ 10% at the 90% confidence level at the system and class peak, and this has become more or less a standard for load research studies.

Accuracy is a function of the sample size and the population variance. In some cases the accuracy requirement will be fixed for you, so you will need to select as large a sample as necessary to meet the requirement, plus some alternates to compensate for non-participants. In other cases, economy might be more important than a high degree of accuracy, meaning a smaller sample size might be acceptable. The Sampling Package allows you to set the desired accuracy and the system will calculate the number of points required; or you can specify a fixed or minimum number of points.

- **What measure of variance will be used to determine sample size?**

Customer-to-customer variation is the basic determinant of sample size. You can estimate this variance by analyzing either: 1) data from a prior study that measured the characteristics of interest for a similar population; or 2) a proxy variable known for your entire target population.

If data from a prior study is available, the first approach is preferable. It typically provides a more reliable variance estimate, thereby improving accuracy while minimizing the number of sample points required. If your utility has not conducted applicable studies in the past, you may be able to borrow data from another utility that has studied a population with similar characteristics and has a similar climate.

If no such data can be found, your other option is to use a proxy variable that is closely correlated to the characteristic of interest. For instance, if you are designing a sample to estimate peak winter demand, you may have to use a proxy from the Customer Information System, such as average January energy.

- **What are the restrictions?**

Of course, you must always keep any budget and time constraints in mind when planning your sample. For example, if you have unlimited budgets and time, you might use a large number of sample points. However, if schedule and cost are issues, you'll need to take care to design a sample that uses only the minimum number of points necessary to meet the desired accuracy.

Once you have answered the preceding questions, you are ready to select a sampling technique (**Chapter 2**: **Selecting a Sample Design Technique and Establishing a Plan**).

Chapter 2

Selecting a Sample Design Technique and Establishing a Plan

After you have established your sample design criteria, you are ready to select a sample design technique and possibly establish a game plan for testing alternative design scenarios.

What Techniques Are Available?

The Oracle Utilities Load Analysis Sampling Package supports three approaches to sampling:

- Non-Stratified Simple Sampling (including Systematic and Centered Systematic)
- Stratified Single Dimensional Sampling
- Stratified Multidimensional Sampling.

These methodologies, along with their relative advantages and disadvantages for different situations, are described below and summarized in Table 2-1.

Table 2-1: A Comparison of Sample Design Techniques

	NON-STRATIFIED SIMPLE SAMPLE	STRATIFIED SINGLE-DIMENSIONAL SAMPLE	STRATIFIED MULTIDIMENSIONAL SAMPLE
Description	Every customer has equal probability of being chosen	Divides population into groups, i.e., STRATA. Models population for the dimension variable e.g., peak month demand	Divides population into cells based on more than 1 criteria. Creates more than 1 model for the population
PROS	Easier to perform. Provides flexibility for future analysis	Saves $. Sample sizes tend to be smaller requiring fewer recorders	May eliminate STRATA migration. Includes characteristics variables
CONS	Sample size larger. Still must examine periods of desired accuracy (e.g., summer, winter)	May provide inadequate estimates for other periods. STRATA migration	More complex to perform. Too many design variables and STRATA make sample unwieldy

Non-Stratified, Simple, Random Sampling

With this technique, individual customers are selected from the population at random. Every customer has equal probability of being chosen.

Advantages of this approach are: 1) easy to perform, 2) provides maximum flexibility for future analyses not anticipated during the sample design phase, and
3) avoids the problem of sample migration in dynamic populations (explained further below).

However, simple random samples are generally less efficient than stratified designs; that is, they require more sample points to meet a specified accuracy level.

Stratified, Random, Single-Dimensional Sampling

In this approach, customers with similar characteristics are grouped together into non-overlapping, homogeneous groups called "strata," and individual samples are selected from each stratum.

The strata are defined according to a user-specified demographic or usage variable called the "design variable." For continuous variables such as usage, the Dalenius-Hodges rule is used to define the strata boundaries. Then, the Neyman allocation procedure is used to determine the optimum sample size for each stratum. (In Neyman allocation, the sample size for each stratum is determined according to its population proportion and the standard deviation. Data from prior load research studies, if available, may also be used to determine the mean and the standard deviation.) A simple random sample is then selected from each stratum.

Because customer-to-customer variation is the basic determinant of sample size (the more the variation, the larger the sample), fewer sampling units need to be selected from a population that has been stratified into homogeneous groups than if the units were merely selected from the entire population at random. In other words, because the variation within a stratum is less than for the entire population, fewer sample points are required to obtain the same accuracy level.

Stratification is a good choice when you need to economize with a smaller sample size, yet maintain a specified level of accuracy. It is also useful when you need data for specific demographic sets within the population (types of business, location, etc.). However, stratification has some aspects which may make it inappropriate for certain situations, i.e., since not all customers have the same chance of being selected, the sample may not be as flexible. Therefore, if you wish to use the sample to perform analyses and answer questions not anticipated in the original design, you may have to employ Domains Analysis to ensure that original sample weights are taken into consideration.

Also, over time, some customers will change their characteristics and will migrate out of their strata. However, the strata assignments must remain fixed throughout the analysis period. For that reason, samples must be replaced periodically to keep them up to date.

Stratified, Random, Multidimensional Sampling

With this approach, the population is stratified by at least two, but no more than seven design variables. The same procedure described above for single dimension stratification is applied for each dimension in the design. The results are then combined into cells reflecting the overlap of the dimensions (see Figure 2-1). A simple random sample is then selected from each cell.

It is possible to use a combination of usage variables, or even a combination of usage and demographic variables in a multidimensional design. E.g, a commercial population might be stratified by seasonal demand and rate code. Another scenario might involve stratification by annual energy, geographic code, and SIC Code.

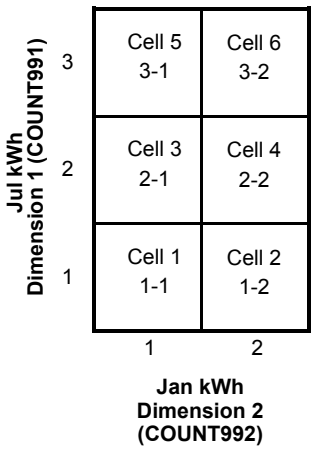


Figure 2-1 Possible Sample Design

The multidimensional stratified approach is useful when you need to satisfy multiple objectives with a single sample, or when there are obvious demographic divisions within the population. However, multidimensional samples are more difficult and time-consuming to perform. Also, more sample points are required to meet the desired accuracy for all levels.

You must take care to limit the number of variables since a large number of strata can quickly lead to an unwieldy design. For example, if you specify three dimensions with four strata in each, you have 64 cells.

In summary, each design methodology offers advantages for different types of situations. The methodology you select for your sample will depend upon the requirements and limitations of your specific project.

Establishing a Plan

Sampling is as much an art as a science, and can involve a lot of experimentation and discovery. While the Sampling Package performs much of the work automatically (such as the calculations), you must still rely upon your experience and knowledge of the population to make a variety of decisions throughout the process (choice of design variables, sampling technique, population definition, accuracy requirements, number of strata, etc.). You will need to experiment with several alternative scenarios before you arrive at the optimal sample for your purposes.

Fortunately, the Oracle Utilities Load Analysis Sampling Package provides a number of options and avenues that make it easy to try out and compare different designs. However, before you sit down at the terminal, you may find it helpful to sketch out a few alternative scenarios for testing with the system.

Table 2-2 illustrates a number of sample design scenarios one Oracle Utilities Load Analysis user company identified at the beginning of its sampling project.

Table 2-2: Sample Design Scenarios for a Load Research Study

Sample design scenarios one Oracle Utilities Load Analysis user company identified at the beginning of its sampling project.

The following are scenarios being used in the new sample designs for three rate classes: **Scenario #1** Rate class, revenue code, summer peak month, and winter peak with all dimensions having 2 to 4 strata. **Scenario #2**

Table 2-2: Sample Design Scenarios for a Load Research Study

Sample design scenarios one Oracle Utilities Load Analysis user company identified at the beginning of its sampling project.

Rate class, revenue code, and the average of the 4 summer months and winter peak months with all dimensions having 2 to 4 strata.

Scenario #3

Rate class, revenue code, winter and summer peak months with high and low usage.

Scenario #4

Use hours use from the old sample to design the new sample using the same criteria as scenarios 1 through 3.

Scenario #5

Use the old sample using post stratification to derive statistics for use in the new sample design for scenarios 1 through 3.

Scenario #6

If time permits, include the SIC as another dimension in scenarios 1 through 3 without any strata.

The objective is a design accuracy of ± 5% at the 95% confidence level, the minimum is ±10% at the 90% confidence level. The Oracle Utilities Load Analysis Sampling Package will be used to design the samples, select the customers to be used in the sample, and validate the sample.

Chapter 3

Creating the Population Data File and Record Definition File

After you have clearly defined your objectives for the sample, and have developed a "game plan" for the design process, you will need to establish three important input files:

- **Population Data File (PDF)** — contains one record for each member of the target population. You will design and draw your sample from this file.

  The PDF is typically extracted by someone in your Information Systems Department from the utility's Customer Information System (CIS) or Billing System database according to specifications you provide.

- **Record Definition File** — defines the record layout of the Population Data File so that other programs in the sampling process can read it. You will create this required file using the Record Definition Program (B110).

This chapter explains how to create these *required* files.

The Population Data File

The Population Data File is the primary input file to the Oracle Utilities Load Analysis Sampling Package programs, so its contents must be carefully defined at the beginning of the project. This file is the main source of data for population analysis, stratification, and sample selection.

The Population Data File is the primary input file to the Oracle Utilities Load Analysis Sampling Package programs, so its contents must be carefully defined at the beginning of the project. This file is the main source of data for population analysis, stratification, and sample selection.

Typically, you will first define the contents of the PDF, then ask an individual in your Information Systems Department to prepare the file for you. You can use the forms provided at the end of this chapter to convey your request. You may wish to print out or make a photocopy of these forms, complete them with the specifics of your project, and give them to the appropriate individual. The forms also define general format requirements and limitations from a data processing point of view. (If you are creating the PDF yourself, these forms can serve as a useful mechanism for organizing and recording your thoughts.)

Note About Using Existing Population Data Files

The Sampling Package has been designed to enable you to take advantage of existing files. If you already have a Population Data File that contains all of the data you require for your sampling project, you can use the Record Definition File to define the layout of that PDF to the Sampling System. The programs can then locate pertinent data and skip over irrelevant fields.

The Population Data File Should Contain:

- A unique customer number or account number
- Selection (optional) & Design variables
- Any additional data for reporting purposes such as name, address telephone number.
- A total of 8 blank spaces at the end of each record.
    - These are blank spaces in which the Sampling Programs will fill in during the design and selection process.

Population Data File Layout:

- Must be fixed length: Each record must to contain the same # of characters.
- If numeric fields with decimal values must be fixed decimal.

Here is a brief list of the steps you will follow to create a PDF. Each of these steps is described in detail on the following pages.

SUMMARY
CREATING THE POPULATION DATA FILE (PDF)

1. Carefully define the contents of the PDF.
2. Fill out the "Population Data File Request" form. Give a copy of that completed form, along with the "Guidelines for Population Data File (PDF) Creation" and a blank copy of the "Population Data File Layout" form to the person creating the PDF.

Step 1: Carefully define the contents of the PDF

Following are some guidelines you should keep in mind when defining the contents of the PDF.

Content

There must be one PDF record for every customer in your target population. Depending upon your sample requirements, this may mean just a specified group of customers (such as a rate class), or every customer served by the utility. The PDF may contain anywhere from several thousand to several million records. Keep in mind that a large file containing many customer groups will give you more flexibility, but will also require more processing time.

> **Note:** Throughout this manual, we refer to the sampling unit as a "customer." Your definition of a sampling unit will be more specific — for example, it might be a customer billing account or a customer premise. Other possible sampling units might include transformers, whole office buildings, or even individual appliances.

It is critically important that *every* record in the PDF contains *all* of the information you intend to use in the sample design, selection, and reporting process (with the exception of data from a prior load research study — see **Chapter 5**: **Stratifying the Population and Determining Sample Size for a Single Dimension** and **Chapter 7**: **Recalculating Population Statistics for a Multidimensional Design Using Prior Load Research Data**). There is no limit to the amount of information you can have in a PDF record (however, only 512 variables can be described by the Record Definition Program, which is explained further in the next section). At a minimum, ***be sure each record contains the following:***

- A unique customer identifier or account number.
- Selection variables (see explanation below).
- Design variables (see explanation below).
- Any additional data you intend to use for reporting purposes, such as customer name, address, and telephone number.
- Four bytes reserved for strata number and four bytes for random number. These are blank spaces that the Sampling programs will fill in during the design and selection process. These fields can be located anywhere in the record, but we recommend that you put them at the end.

The file can contain fields not used by the sampling programs, since you can direct the software to "skip over" irrelevant fields. You will also have the opportunity to "overlay" fields (use selected fields for multiple purposes). These two features are explained under "How to Create the Record Definition File" on page 3-8.

About Selection and Design Variables: *Selection variables* are data you will use to identify the target population (required only if your target population is a subset of the PDF). *Design variables* are data you will use to segment or stratify the population. Either type of variable can be any demographic or historical usage data available for the entire population. When planning the PDF, you may ask for several alternatives, so you can experiment with different versions later in the design process. *Whatever data you choose must be available for every customer and must be defined and handled in a consistent manner.*

For example, let's say you intend to design separate samples for residential and agricultural classes. You might request that the Information Systems Department create two files, one for each class; or, to shorten the lead time for obtaining the data, you might request one file containing both classes. In the second case, you would need to be sure that a class identifier was contained in every record.

Or, let's use a more complicated example. Suppose you intend to design a multidimensional stratified sample that will provide accurate estimates of summer and winter peak demand for

residential customers in the Northern and Southern service territories. In this case, you will need a geographic code to divide the population into the specified subgroups, and you will need usage data to use as stratification variables. Since you probably don't have current summer and winter demand data for the population, you will need to choose proxy variables that are available in your billing files and are closely correlated with the target demands. Since at this point you may not know the variables with the closest correlation, you might ask for several alternatives; such as billed energy for the months of January and July, or average monthly energy for December-January-February and for June-July-August. (You will have the opportunity to evaluate the alternatives later in the Population Analysis phase.)

This brings us to another very important point about design variables. ***If a design variable requires calculation, it must be performed before input into the PDF***. Data in the Customer Information System — especially usage data — often requires manipulation before it can be used in the sample design process. For example, you may wish to use average seasonal energy as a design variable, but the CIS system contains only energy use by billing cycles. Since the Sampling programs do not perform calculations on the variables, the Information Systems Department (or whoever is preparing the PDF) would have to use an external program to calculate average seasonal energy from the billing data before inputting the values into the PDF.

Examples of common adjustments to usage variables that would have to be made outside the Sampling programs include:

- Normalizing from billing cycle to calendar periods
- Combining multiple meter readings
- Prorating bimonthly, incomplete or partial readings
- Summing or averaging for annualized usage.

Other Considerations

It is also important to anticipate situations that require interpretation, and give the Information Systems Department guidelines for handling them. For example, customers with zero usage are often an issue. New customers will probably have zero usage for part of a year. In this case it is acceptable to have zero usage, and that fact should be incorporated into the sample design. At other times, however, zero usage might indicate billing file problems, which obviously should not be incorporated into the design. Each utility is unique; you will have to develop policies for your particular situation.

Aside from new accounts/customers, other issues you may need to address include:

- Master metered accounts
- Seasonal customers
- Tenant changes
- Rate class changes.

In summary, all data you specify for PDF must be:

- available for every customer in the population
- computed before input to the PDF
- specified in a consistent and logical manner, anticipating ambiguous situations.

Step 2: Fill out the "Population Data File Request" form

Fill out the "Population Data File Request" form, and Give a copy of that completed form, along with a blank copy of the "Population Data File Layout" form in this chapter, to the person creating the PDF.

Once you have established your requirements for the PDF, record them in the "Population Data File Request" form. Give the completed form and the "Guidelines for Population Data File (PDF) Creation" to the person actually creating the file. (***Note:*** These materials are provided at the end of this chapter. If you photocopy the forms, be sure to copy both sides.)

Your next task (explained in the following section) is to create the Record Definition File, which defines the layout of the PDF so that the Sampling programs can read it. Before you can construct this file, you'll need to get a description of the contents and format of the PDF from the person who created it. To facilitate this process, we've provided a "Population Data File Layout" form, which you can photocopy and give to that individual to record the setup of the PDF for you. This file must have the .PDF extension. Once your PDF file created, your next step is to import it to a new sampling project.

Sampling Projects

Sampling Projects allow you to group all related input files together so that as you progress through the various steps in Sampling, the only files available in your drop selections are ones related to your current project. In addition, outputs created by Sampling Programs that are used at input in a subsequent step is automatically copied to the current Sampling Project and made available for use in the subsequent step.

File Structure Layout

All Sampling Project files and folders will be kept and managed under the COMMON\DATA\SAMPLING directory. The structure will be as follows:

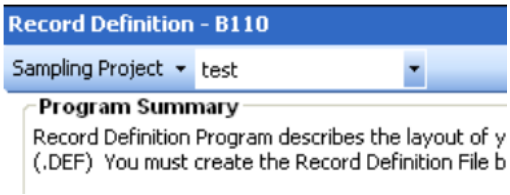
```
[COMMON]
  [DATA]
    [SAMPLING]

[PROJECT NAME]
      [B110]    <- Program Specific Folders
        *.ctl
        *.fdf
      [B210]
      [..etc]
      *.PDF     <- Common Sampling Files
      *.DEF
      *.SPF
      *.RCF
```

- **Program Specific Folders** – Contain input files that are program specific, usually these are CTL, FDF files.
- **Common Sampling Files** – Contain input files that are not program specific, usually these are PDF, DEF, SPF, RCF files.

Creating a New Sampling Project

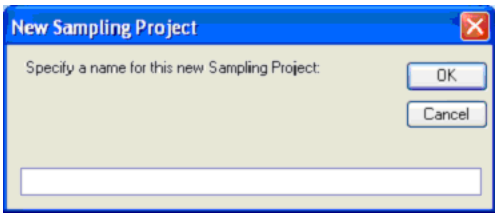
To start a new Sampling Project, navigate to the Sampling Program, B110 Record Definition. At the top of the submit panel, the Sampling Project Toolbar will be visible:



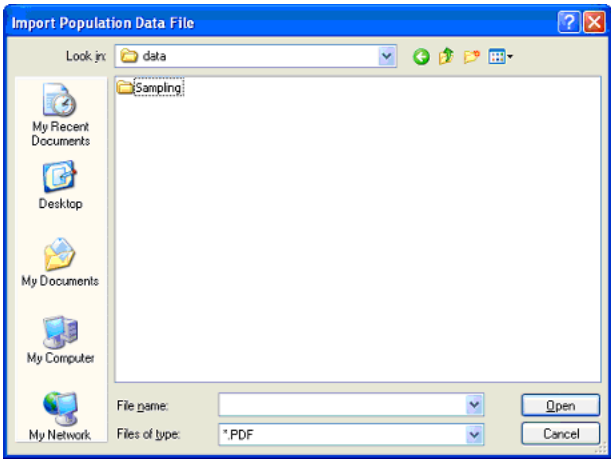
The toolbar is broken down into two components:

1. **Sampling Project Drop Down Menu** – The Sampling Project Menu allows you to create, delete, and manage projects. Available menu items are as follows:
   - New Sampling Project – Creates a new, empty Sampling Project Folder.
   - Import Population Data File – Once a project has been created, you can import an existing PDF file for use with the current project. Once imported, the PDF should be available for selection in the drop downs for the project.
   - Delete Sampling Project – Deletes an existing Sampling Project and all files associated with it.
2. **Current Sampling Project** – A combo box drop down that lists the currently selected project.

Click on **Sampling Project** -> **New Sampling Project** to start your new Sampling Project. You will be prompted to specify a name for this project:



Once you have created your new Sampling Project, you are ready to import your Population Data File (PDF) into this project. To import a PDF into the current Sampling Project, click on **Sampling Project** -> **Import Population Data File**. You will be prompted to select the PDF file to import:



Once you have successfully imported your PDF to your new Sampling Project, you are now ready to define the layout of your Population Data File. You will do this by creating the Record Definition File in the next step.

How To Create the Record Definition File

The Record Definition File is an internal table of variable names and datatypes that defines the layout of the PDF. It is stored as a binary file in your sampling project's directory. You **must** create the Record Definition File before the PDF can be processed by the Population Analysis or Sample Selection programs.

You will use the Record Definition Program to create the Record Definition File (.DEF). The.DEF extension must be included (which the GUI will do automatically). Here is a brief list of the steps you will follow. Each of these steps is described in detail on the following pages.

SUMMARY
CREATING THE RECORD DEFINITION FILE
USING THE RECORD DEFINITION PROGRAM (B110)

1. Create the Record Definition Control File (TGB12A.CTL) — a definition of the physical record format of the Population Data File.
2. Submit the job (Record Definition Program — B110).

Step 1: Create the Record Definition Control File (TGB12A.CTL)

The Record Definition Control File (.DEF) describes the physical record format of the Population Data File (.PDF) so that other Sampling programs can read it. You can create this file manually via the composer.

Record Definition Control File

As shown in the sample in Figure 3-1, you use **field definitions** to define the variable name, datatype, and optional comments for each field in the PDF record. For instance, the first 20 characters in the example below are defined as CUSTID, which consists of RATE, ROUTE, FOLIO, and TENANT.

The DELIMITER parameter specifies the character that is used to separate fields in the population data file. The parameter supports either a comma or a semicolon. The semicolon is the default value.

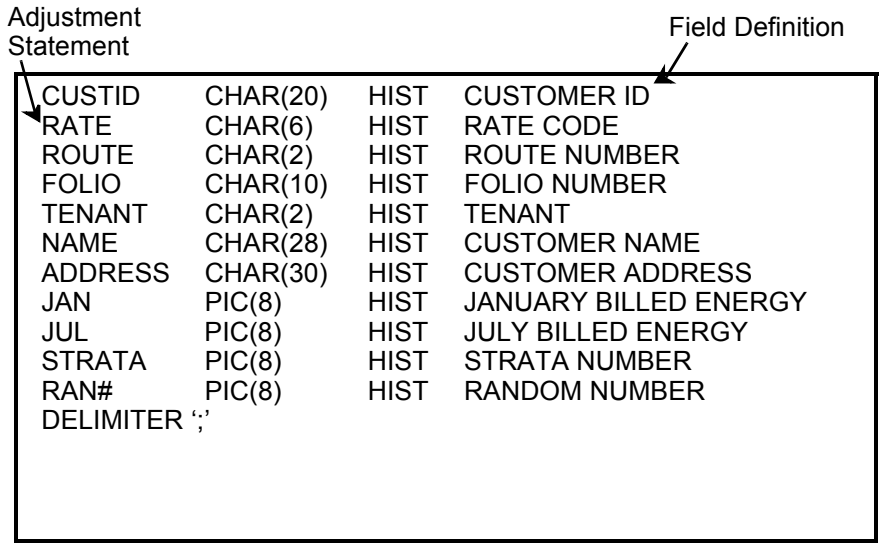


Figure 3-1 Sample Record Definition Control File

Field Definitions Statements

Use the format shown below to create each field definition. You can input just one field definition statement per line. A maximum of 512 field definition statements may be specified in a Record Definition Control File.

variable name	datatype	HIST comment

- **variable name** — a unique one- to eight-character identifier. You can use any printable character except commas, parentheses, and apostrophes. It is recommended that you use easy-to-remember and logical names, since you will have to use them throughout the sampling process.

Important: to maximize processing performance, try to set it up so that the first, second, and last characters of each name create a unique identifier for all fields in your Control File. For example, avoid having two field names like ACCOUNT# and ACTION# in your file, since both would be read AC#. The program will accept them, but processing time will be extended.

- **datatype** — use this parameter to specify the type of data contained in the field. (This information should be available from the person who created the file.) The possible types are:

  CHAR(n) — Character strings of a length specified by "n". ("n" bytes)

  PICTURE(n) or PIC(n) — Use Picture data to define numeric fields with field's length specified by "n", which must be a positive integer or positive fixed decimal.

  Note: The only human-readable datatypes are CHAR and PICTURE data.

- **HIST** — you must input "HIST" to separate datatype and comments. (It is not required if you are not using comments.)

- **comments** — optional descriptor you can use to record an explanation of the field's contents, or other notes you feel would be helpful to you and others. This field is ignored by the Record Definition Program.

  Note: You must include these two field statements in your Control File (with comments if desired):

  STRATA PIC(8)

  RAN# PIC(8)

  Both are required fields that will be used by the Sampling programs during processing. STRATA is used for strata assignments; and RAN# for random number assignments. They can be placed anywhere in the file; at the end is recommended. Each requires 4 bytes, for a total of 8 bytes for the two fields.

Step 2: Run the Record Definition Program

Use the Submit panel to run the record definition program.

Note: The buffer size for the sampling program is specified in the CSLSTAR.GLB configuration file. For more information, see **CSLSTAR.GLB (Run Time)** on page 1-6 of the *Oracle Utilities Load Analysis Configuration Guide*.

Guidelines for Population Data File (PDF) Creation

The following summarizes requirements and restrictions for creating a Population Data File (PDF) for use with the Oracle Utilities Sampling Package. The file must contain the following information:

- A unique customer identifier or account number for each record.
- All adjustments must be precomputed during creation of the PDF for use by Sampling. Common adjustments include combining multiple meter readings, prorating bimonthly, incomplete or partial readings, and summing or averaging for annualized usage.

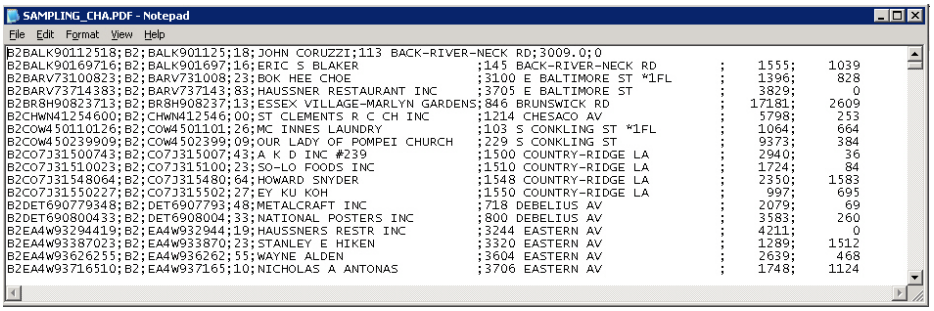
The Sampling Package can accommodate an existing Population Data File not specifically designed for this application as long as the file contains all required data.

The file is unrestricted in terms of record length or number of records. However, it is delimited by the character specified in the DELIMITED command in the record definition control file. It is essentially unrestricted with respect to field placement, although fullword alignment will reduce execution time for numeric values. The following datatypes are allowed:

- **Character** — 1 to 256 bytes, left justified. Any string of alphanumeric characters up to 256 bytes.
- **Picture** — n bytes, as specified in the Record Definition File
  - Readable numeric data.
  - May contain leading and/or trailing blanks.
  - May contain up to four decimal points.
  - Numeric fields with decimal values must be fixed decimal.

    Note: Fields that contain a special character, such as a semicolon or double quotes, must be enclosed in double qoutes. If a field contains a double qoute character, it is escaped by placing another double quote character next to it.

Sample Population Data File:



Population Data File Request

To: John Smith (MIS) Date: August 16,1992

From: Jane Doe (Load Research) Date file required: September 16, 1992

Phone/Ext: x555 Send to: Jane Doe, Mail Stop 301

Please prepare a file of customer information according to the specifications below and the requirements printed on the back of this form. The file will be input to the Load Analysis Sampling Package and analyzed to create a statistically-reliable sample of the customer population. Once you have created the file, ***please give me a written description of the data file layout*** (variable name, starting position, length in bytes, and datatype for each field). Only the requested fields need be described. I have provided a "Population Data File Layout" form for you to fill out to help facilitate this process.

Additional comments:

Customer information for January and July 1901.

Do not include customers with zero usage for the specified periods.

Variable Name*	Description	Required Computations	Comments
NAME	customer name	none	
ADDRESS	customer address	none	
CUSTID	customer id		unique identifier
RATE	rate code	none	
ROUTE	route number	none	
FOLIO	folio number	none	
JAN	January billed energy	reconcile rebills and	will be used for analysis — do
		60 day bills	not assign character datatype
JUL	July billed energy	same as above	same as above
STRATA	Strata number	none	reserve field — 4 bytes — required
RAN#	Random number	none	reserve field — 4 bytes — required

*maximum of 8 characters

Page 1 of 1

Population Data File Request

To: _____ Date: _____

From: _____ Date file required: _____

Phone/Ext: _____ Send to: _____

Please prepare a file of customer information according to the specifications below and the requirements printed on the back of this form. The file will be input to the Load Analysis Sampling Package and analyzed to create a statistically-reliable sample of the customer population. Once you have created the file, ***please give me a written description of the data file layout*** (variable name, starting position, length in bytes, and datatype for each field). Only the requested fields need be described. I have provided a "Population Data File Layout" form for you to fill out to help facilitate this process.

Additional comments:

Variable Name*	Description	Required Computations	Comments

*maximum of 8 characters

Population Data File Layout

This form should be used to describe the layout of the Population Data File (.PDF). Please list the fields in the order they occur on the file. A sample is printed on the back of this form for your reference.

Prepared by: _____ For: _____

Phone/Ext: _____ Date: _____

File name: _____

Media: _____

Additional comments:

_____

_____

_____

Variable Name*	Starting Position	Length (bytes)	Datatype	Decimal Placement (n,d) for packed data	Description

*maximum of 8 characters

Population Data File Layout

This form should be used to describe the layout of the Population Data File (.PDF). Please list the fields in the order they occur on the file. A sample is printed on the back of this form for your reference.

Prepared by: John Smith (MIS) For: Jane Doe (Load Research Analyst)

Phone/Ext: X766 Date: September 10, 2001

File name: CUSTINFO

Media: Tape

Additional comments:

Variable Name*	Starting Position	Length (bytes)	Datatype	Decimal Placement (n,d) for packed data	Description
CUSTID	1	20	CHAR(20)		breaks down into rate, route, folio & tenant
RATE	1	6	CHAR(6)		
ROUTE	7	2	CHAR(2)		
FOLIO	9	10	CHAR(10)		
TENANT	19	2	CHAR(2)		
NAME	21	28	CHAR(28)		
ADDRESS	49	30	CHAR(30)		
JAN	203	4	INTEGER		
JUL	251	4	INTEGER		
STRATA	307	4	INTEGER		
RAN#	311	4	REAL		

*maximum of 8 characters

Chapter 4

Analyzing the Population Frequency Distribution

The first task in the Sample Design Phase will be to use the **Population Analysis Program** to create a frequency distribution of your target population based on a usage variable. You will specify the target population, usage variable, and frequency interval ranges; the Population Analysis Program will sort and tally the customers in the Population Data File accordingly. The resulting Frequency Distribution File (.FDF) will become an input to the Sample Design Program and used as the basis for stratification.

You will typically perform this task no matter which of the three sample design methodologies you have selected. If you are creating a multidimensional stratified design, you will perform this task for each usage variable in the design. Even if you are ultimately creating a single dimensional stratified or simple random design, you may need to perform this task a number of times to arrive at an optimal interval scheme, and/or to evaluate alternative usage variables.

Creating the Frequency Distribution File (.FDF)

Following is a brief list of the steps you will follow when creating the Frequency Distribution File (.FDF) using the Population Analysis Program. These steps are described in detail on the following pages.

SUMMARY
CREATING THE FREQUENCY DISTRIBUTION FILE USING THE POPULATION ANALYSIS PROGRAM (B210)

1. Verify or create required input files:
   - Population Data File (.PDF)
   - Record Definition File (.DEF)
   - Analysis Control File (TGB22A.CTL) — population selection and interval definitions.
2. Submit the job (Population Analysis Program — B210).
3. Check outputs.
   - Population Analysis Execution Log (TGB220-01)
   - Frequency Distribution File (.FDF) — will be used as the Control File (TGB31A.CTL) for B310, Single Dimensional Sample Design in **Chapter 5**: **Stratifying the Population and Determining Sample Size for a Single Dimension**.
   - The Frequency Distribution Graph

Step 1A: Verify that the Population Data File is Available

Verify that the Population Data File (.PDF) has been properly set up and is available. It should be available for selection in the submit panel under the Population Data file dropdown. If you have not already created the Population Data File (.PDF), see **Chapter 3**: **Creating the Population Data File and Record Definition File**.

Step 1B: Verify that the Record Definition File is Available

Verify that the Record Definition File (.DEF) has been properly set up. It should be available for selection in the Record Definition file dropdown under "B210 Population Analysis" submit panel. If you have not already created the Record Definition File (.DEF), see **Chapter 3**: **Creating the Population Data File and Record Definition File**.

Step 1C: Create the Analysis Control File

You use the Analysis Control File to:

- Select the target population (this is necessary only if the target population is a subset of the Population Data File. If every member of the PDF is eligible for the sample, you do not need to do this).
- Specify the usage intervals into which the program will classify the target population.

When you create the Analysis Control File, your objective is to define an interval scheme that models the actual distribution of the population. To that end, you may vary interval sizes as needed, depending upon where the population is concentrated. For example, residential populations may require many small intervals at the lower end of energy usage and fewer larger intervals as usage increases.

The important thing to keep in mind is that the Sample Design Program (next chapter) will use your interval breakpoints as the basis for stratification. In other words, *the Sample Design Program can assign a stratum boundary only where an interval breakpoint already exists.* The more intervals you have, the more choices the Sample Design Program has for assigning strata breakpoints. Also, small intervals are necessary to avoid inaccuracies caused by a skewed population. The program uses the midpoint of block usage to approximate the mean. As long as the intervals are small, this is a valid approach.

Therefore, your objective is to "break up the lumps" — that is, divide intervals with a large population into small intervals with fewer members. And don't combine intervals that seem to have too few members. You can't have too many intervals, but you can have too few. A good rule of thumb is to have at least twenty to thirty interval breakpoints in your Analysis Control File.

Determining the number and size of usage intervals can be an interactive process, in which you define an interval scheme in the Analysis Control File, run the program to test it, then modify the scheme and rerun the program. You may have to do this process a number of times until you arrive at a satisfactory distribution, especially if you are unfamiliar with the characteristics of the population. To help simplify the process, it's recommended that you use the Analysis Control File provided by Oracle Utilities Load Analysis with just slight modification for the first pass. Or, you might use an Analysis Control File developed for a previous, similar study. (By starting with an existing file, you avoid the necessity of typing an entire file from scratch.) The initial run will show you how the population is distributed. You can then modify the Analysis Control File, adding finer intervals to those areas where the population is concentrated.

Creating the Population Analysis Control File

The Population Analysis Control File determines how the Population Analysis Program selects, sorts, and counts customers in the Population Data File. Creation of the Population Frequency Distribution is determined by the following Control File commands:

SELect	field = value
USAge	usageVariable
ENDpoint	endpoint1, endpoint2, …. endpointN
SIZe	intervalSize1, intervalSize2, …. intervalSizeN
CEIling	ceiling

SELect field = value

SELECT (optional) - SELECT is a test statement for defining eligible customers. For example, to process records whose RATE code is 'SGS' only: SELECT RATE = 'SGS'

USAge usageVariable

USAGE - Defines the usage, or design variable. Specify the name of the usage variable you would like to use. For example:

USAGE JAN

ENDpoint endpoint1, endpoint2, …. endpointN

SIZe intervalSize1, intervalSize2, …. intervalSizeN

ENDPOINT – A comma separated list of endpoints. Each endpoint should have a matching interval size (below). Endpoints, in conjunction with interval size, determine the list of interval boundaries in which to count the number of customers falling into each interval.

SIZE – Specify the interval size of each endpoint provided and they need to be comma separated. Interval size determines the number of breakpoints to allocate up to the endpoint.

For example:

```
ENDPOINT 100, 500
SIZE     25, 100
```

The above example will create interval boundaries each with an interval size of 25 up to the first the first endpoint (100). Then for usage between 100 and 500, create boundaries with interval sizes of 100 each. The resulting list of interval boundaries from which to count customers falling into would be: 25, 50, 100, 200, 300, 400, 500.

CEIling ceiling

CEILING - Provide largest possible usage value in your population. This value should be larger than the last endpoint.

Sample Control File

```
USAGE     JAN
ENDPOINT  50, 100, 250, 1000, 1200, 1500, 2000, 10000, 50000, 100000
SIZE      5, 25, 50, 250, 100, 150, 500, 1000, 40000, 50000
CEILING   123000
```

About the User Language (Optional)

You can optionally construct the Analysis Control File using a powerful "User Language." If you already have constructed your analysis control file, you can skip this section. The "User Language" allows for more customization in the logic used to create your frequency distribution file, but keep in mind it is also more complex.

Essentially, the Population Analysis Program selects, sorts, and counts customers in the Population Data File according to logic you define with the User Language. Because this language is very powerful and has many uses, it is also somewhat complex. To use this language fully, you will need to read the description provided in Appendix A carefully. However, for your convenience, we have described here just the statements you need to create or modify a typical Analysis Control File.

See the template and examples shown in Figure 4-1 through Figure 4-3.

As you can see, a typical Analysis Control File is made up of a few basic elements:

1. **Comments** — optional notes that do not affect processing. You must enclose any comments between these symbols: /* and */. Do not split comments across lines, unless you begin and end each line with the slash/asterisk symbols. Do not enter the slash/asterisk symbols in columns 1 and 2 of the file; otherwise, you may put comments anywhere in the file.

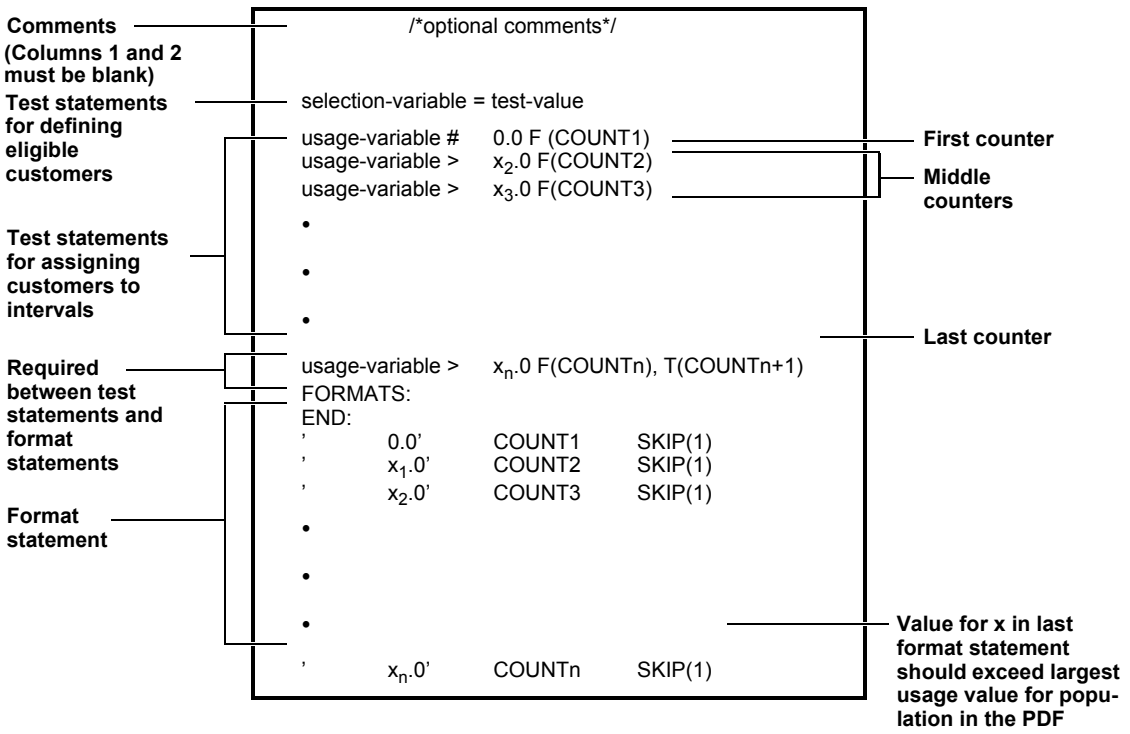


Figure 4-1 Analysis Control File "Template"

Add test statements if target population is a subset of the .PDF file.

Replace JAN with your usage variable (must be a variable name from Record Definition File (.DEF)).

```
JAN     #          0      F(COUNT1)
JAN     >          5      F(COUNT2)
JAN     >         10      F(COUNT3)
JAN     >         15      F(COUNT4)
JAN     >         20      F(COUNT5)
JAN     >         25      F(COUNT6)
JAN     >         30      F(COUNT7)
JAN     >         35      F(COUNT8)
JAN     >         40      F(COUNT9)
JAN     >         45      F(COUNT10)
JAN     >         50      F(COUNT11)
JAN     >         75      F(COUNT12)
JAN     >        100      F(COUNT13)
JAN     >        150      F(COUNT14)
JAN     >        200      F(COUNT15)
JAN     >        250      F(COUNT16)
JAN     >        500      F(COUNT17)
JAN     >        750      F(COUNT18)
JAN     >       1000      F(COUNT19)
JAN     >       1100      F(COUNT20)
JAN     >       1200      F(COUNT21)
JAN     >       1350      F(COUNT22)
JAN     >       1500      F(COUNT23)
JAN     >       2000      F(COUNT24)
JAN     >       3000      F(COUNT25)
JAN     >       4000      F(COUNT26)
JAN     >       5000      F(COUNT27)
JAN     >       6000      F(COUNT28)
JAN     >       7000      F(COUNT29)
JAN     >       8000      F(COUNT30)
JAN     >       9000      F(COUNT31)
JAN     >      10000      F(COUNT32)
JAN     >      50000      F(COUNT33)
JAN     >     100000      F(COUNT34), T(COUNT35)

FORMATS:
END:
'          0.0'     COUNT1    SKIP(1)
'          5.0'     COUNT2    SKIP(1)
'         10.0'     COUNT3    SKIP(1)
'         15.0'     COUNT4    SKIP(1)
'         20.0'     COUNT5    SKIP(1)
'         25.0'     COUNT6    SKIP(1)
'         30.0'     COUNT7    SKIP(1)
'         35.0'     COUNT8    SKIP(1)
'         40.0'     COUNT9    SKIP(1)
'         45.0'     COUNT10   SKIP(1)
'         50.0'     COUNT11   SKIP(1)
'         75.0'     COUNT12   SKIP(1)
'        100.0'     COUNT13   SKIP(1)
'        150.0'     COUNT14   SKIP(1)
'        200.0'     COUNT15   SKIP(1)
'        250.0'     COUNT16   SKIP(1)
'        500.0'     COUNT17   SKIP(1)
'        750.0'     COUNT18   SKIP(1)
'       1000.0'     COUNT19   SKIP(1)
'       1100.0'     COUNT20   SKIP(1)
'       1200.0'     COUNT21   SKIP(1)
'       1350.0'     COUNT22   SKIP(1)
'       1500.0'     COUNT23   SKIP(1)
'       2000.0'     COUNT24   SKIP(1)
'       3000.0'     COUNT25   SKIP(1)
'       4000.0'     COUNT26   SKIP(1)
'       5000.0'     COUNT27   SKIP(1)
'       6000.0'     COUNT28   SKIP(1)
'       7000.0'     COUNT29   SKIP(1)
'       8000.0'     COUNT30   SKIP(1)
'       9000.0'     COUNT31   SKIP(1)
'      10000.0'     COUNT32   SKIP(1)
'      50000.0'     COUNT33   SKIP(1)
'     100000.0'     COUNT34   SKIP(1)
'     123000.0'     COUNT35   SKIP(1)
```

Replace last value with a value that is greater than largest customer usage value in your .PDF file.

Figure 4-2 Example Analysis Control File

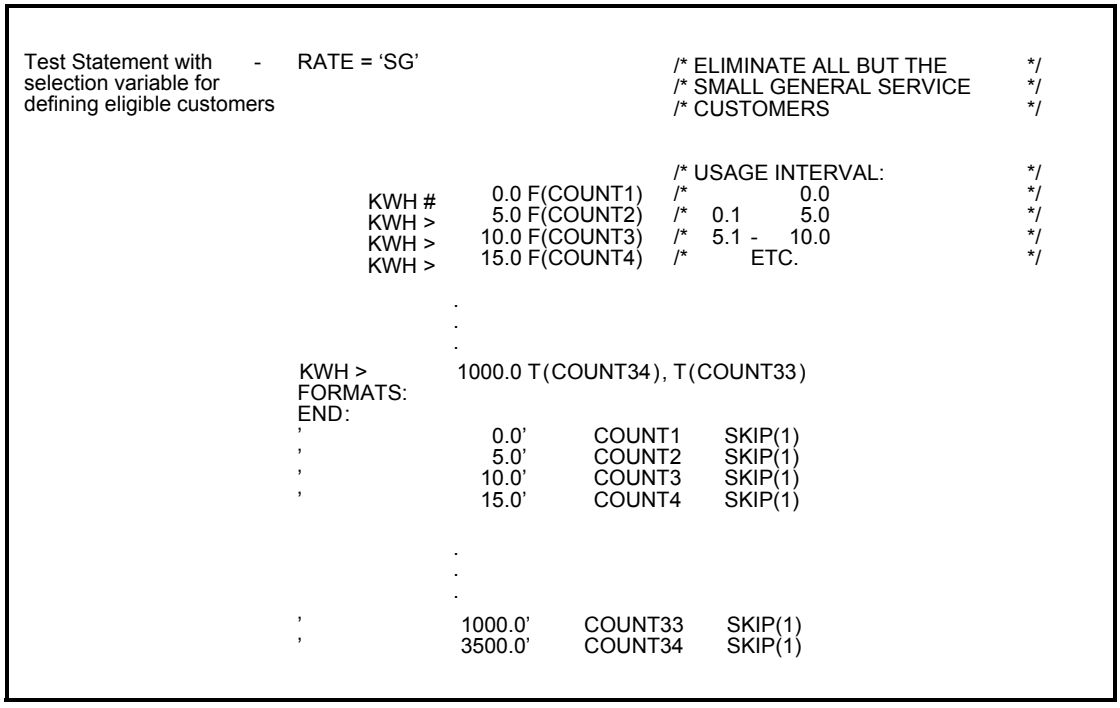


Figure 4-3 Example Analysis Control File with Selection Variable

2. **Test statements** — test statements are used for two tasks in a typical Analysis Control File:1) to select a subpopulation of the PDF for analysis; and 2) to define interval breakpoints and count the number of customers in each interval. **Test statements for selecting eligible customers:** required only if the target population is a subset of the PDF — if you wish to analyze just a particular rate class or demographic group within the PDF, for example. Each test statement compares a single field in a customer's PDF record (called the "selection variable") to your criteria (called the "test value"). If a customer record meets the criteria, it will be accepted and placed in the frequency distribution; if not, it will be passed over. The format is typically:

   `selection-variable = test-value`

   Selection variables must be one of the user-defined variable names in your Record Definition File. For an example, see the sample Record Definition File shown in Figure 3-1. If your facility were using that file, the eligible variable names would be CUSTID, RATE, ROUTE, FOLIO, and so on.

Also, the test value must be a constant or character string that may exist in the PDF records. If the variable is defined as character type data in your Record Definition File and your test value is a character string, you must enclose the test value in single quotes (i.e., 'SG').

If you wish to define the target population using multiple selection variables, or using a relationship other than = (that is, variable equal to test value), see **Appendix A**: **The User Language**.

Test statements for interval boundaries — used to define interval boundaries and count the number of customers falling in each interval. The test statements compare a specified "usage variable" in a customer's PDF record to a series of values you specify. Each value is the upper boundary of a frequency range.

These statements are typically set up in this format:

usage variable > value F(COUNTn)

where:

usage variable is one of the user-defined names in your Record Definition File,

value is the upper boundary of a frequency range, and

n is a sequential number beginning with 1.

Using the logic you construct in the series of statements, the program compares each customer's value for the usage variable to the series of values until it finds the interval in which the customer belongs and increments the count for that interval. More specifically, each statement says to the program, if it is true that a given customer's value for the usage variable is greater than the interval boundary, go to the next statement; but if the customer's value is less than the specified value, activate the counter. Which counter the customer triggers (COUNT1, COUNT2, etc.) determines which interval count is incremented.

The first and last interval test statements require special attention. The first should always be:

usage variable # 0.0 F(COUNT1)

which places all customers with zero usage in the first interval. ***Note:*** It is important to count zero usage customers, but this category *cannot* be input to the Sample Design program since it is not a usage end point.

The last test statement should always be:

usage variable > x_n.0 F(COUNTn), T(COUNTn+1)

which means, if the customer's value for the usage variable is less than the upper boundary specified in the last test statement (x_n), count that customer as belonging in the nth interval; if it is true that the customer's value for the usage variable is greater than x_n, count the customer in the next (that is, last) interval, (n+1).

3. **FORMATS:**
   END:

   You must put these two keywords between the test statements and format statements, just as shown in the examples.

4. **Format statements** — define the layout of the Frequency Distribution File. Format statements are typically set up this way:

   ‘ x_n.0’ COUNTn SKIP(1)

   where **n** is a sequential number beginning with 1.

For each interval test statement in the Analysis Control File, there must be one corresponding format statement. For example, for the counter statement KWH > 5.0 F(COUNT2) you would create the following format statement:

```
‘    5.0’  COUNT2   SKIP(1)
```

where:

- **‘ 5.0’** tells the program to print 5.0 (everything between the single quotes including the blanks).
- **COUNT2** tells the program to print the number of customers counted in the second interval.
- **SKIP(1)** tells the program to skip one line; that is, go to the next line before printing the next count.

Last Format Statement — In addition to the format statements that correspond to each interval test statement, you must include one format statement at the end of the file that identifies the upper boundary of usage variables in the PDF. That is, the value for x in the last format statement

should be greater than the largest customer usage value in the PDF. You may find this value by looking at the population, or you may already know this number from rate blocks.

Remember, your real focus in creating the Analysis Control File is to select the target population and to effectively structure the frequency distribution by selecting a good set of values as interval boundaries. If you start with an existing file and simply specify the target population, the desired usage variable, and last format statement, you can concentrate on the interval boundaries and avoid getting involved in the complexities of structuring the logic. (See Figure 4-2.)

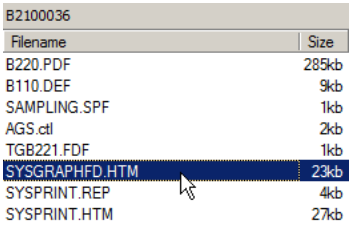
Step 2: Submit the job (B210)

Use the B210 Submit screen, using the previously created Control File, .DEF file, Population Data File (PDF), and the Sampling Parameter file for the selected PDF file.

Frequency Distribution Graph - SYSGRAPHFD

The frequency distribution graphing file is produced from a successful B210 Population Analysis run. This file allows you to graphically view your Frequency Distribution results.

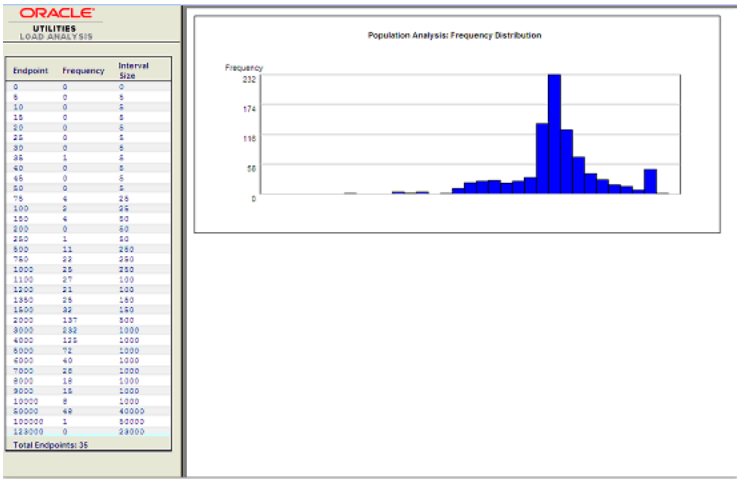
You can view the Frequency Distribution graph by double-clicking on SYSGRAPHFD.HTM file in the lower right of the Results panel portion of the screen:



Sample Frequency Distribution Graphing File.

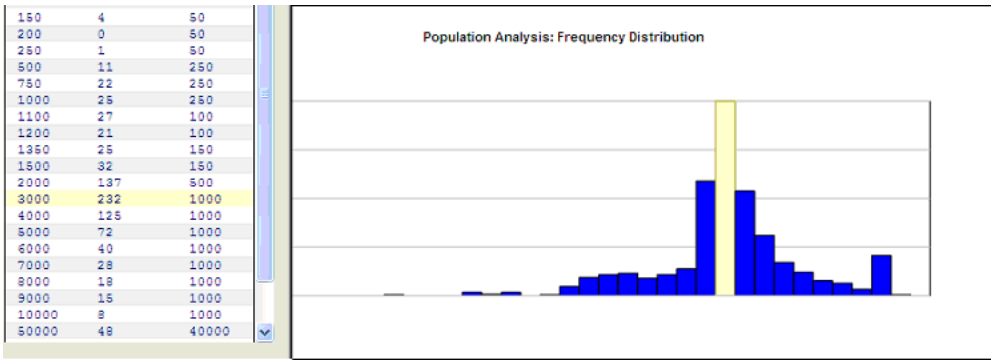
The graphing file is broken down into two components:

1. Frequency Distribution Data Table (left)
2. Frequency Distribution Graph (right)



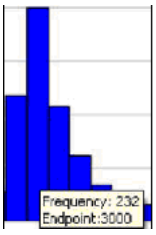
Frequency Distribution Data Table – This contains detail data of all the endpoints and their frequencies. The endpoints are based on what is provided in your Population Analysis Control File. Most likely, you will have to adjust your endpoints and do several runs before you can achieve a frequency distribution that is satisfactory.

Highlighting – Clicking on any values in the data table will highlight the corresponding value in the graph. **Ctrl+Click** can be used to highlight multiple values. Use this to highlight where endpoint(s) are located in the graph.



Frequency Distribution Graph – This is a graph of how your population is distributed based on your population data and the endpoints provided. The shape of your frequency distribution should reflect the demographics of your population. If not, you should go back to your control file, adjust your endpoints, and re-submit the job. A good frequency distribution is important since strata breakpoints and sample sizes will be calculated based on it.

Popup Displays – The line bars in the graph has a mouse over effect. When the mouse pointer is placed over any point in the graph, a yellow popup displays information for that particular point. You can use this to determine the endpoints and frequency values of problem areas in your frequency distribution.



Chapter 5

Stratifying the Population and Determining Sample Size for a Single Dimension

After you have created the Frequency Distribution File (.FDF) and it has been copied to your local DATA directory, you will use the **Single Dimensional Sample Design Program** to define strata boundaries and determine the sample size for a single dimensional stratified design. You will also have the opportunity to use load data from prior studies to improve the quality of your new design.

- Note about using the Single Dimensional Sample Design Program for other types of designs:

  If you are designing a multidimensional stratified sample: you will apply the Frequency Distribution and Single Dimensional Sample Design Programs *once for each usage variable in your design*. However, your objective is to determine strata breakpoints only — you will calculate sample size using another program later in the process.

  If you are designing a simple random sample, you will follow all of the steps described in this chapter, but specify just one strata in the Environment File.

Figure 5-1 illustrates how the Single Dimensional Sample Design Program relates to other programs in the design and selection process, depending upon the type of sample you are creating and whether or not you are utilizing prior load research data.

About Sample Design Program Options

Within the Single Dimensional Sample Design Program itself, you have several options for customizing the design to your specific circumstances. You will use the Environment File to specify them. You may wish to run the program a number of times with different Environment Files to see which options yield the best results. The options include:

- **Strata boundaries**

  You can either supply your own strata breakpoints or apply the Dalenius-Hodges procedure to calculate them. With the Dalenius-Hodges procedure, the program divides the population in the Frequency Distribution File into short intervals. Each interval has frequency f and interval length u. The quantity $\sqrt{uf}$ is summed over all the intervals, and this cumulative $\sqrt{uf}$ is divided by a user-defined number of strata to give the optimum length of each stratum.

  You may supply your own breakpoints if you wish to match your design to a previous study or if you are using rate blocks.

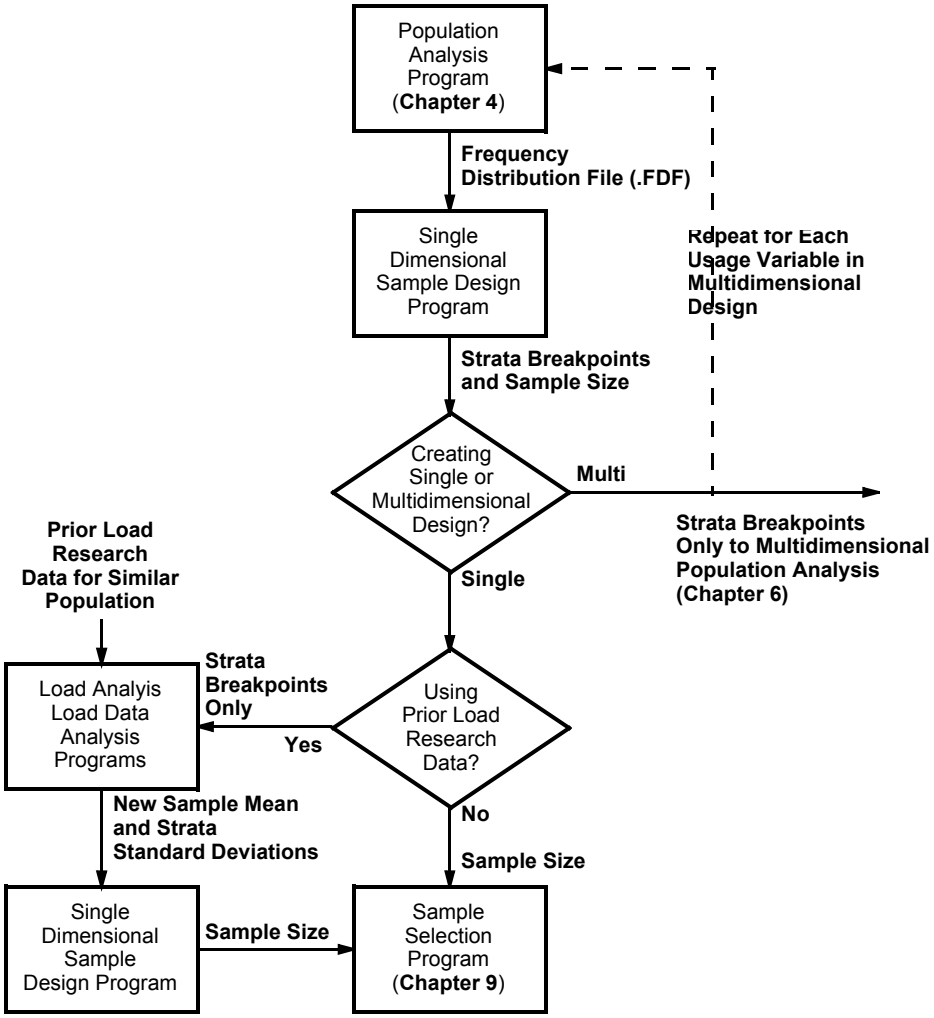


Figure 5-1 Stratifying the Population and Sizing the Sample

- **Sample Size**

  You can either specify a fixed number of sample points, or you can request that the program calculate the optimal sample size based on a user-specified accuracy requirement and number of strata. (By optimal sample size, we mean the smallest number of sample points to meet a specified accuracy.) To find the optimal size, the program uses Neyman allocation to determine the number of sample points for each stratum based upon the percentage of the total population standard deviation represented by the stratum.

Specifying minimum number of sample points per stratum — A new feature has been added to the Sample Design Program, which enables you to specify that a minimum number of sample points be allocated per strata, for either an optimal or fixed sized sample. *For an optimal design,* the program will allocate at least the minimum number of points per stratum, even if it exceeds the number required to meet the desired accuracy. (Of course, the program will not specify a number greater than the stratum population.) *For a fixed size sample*, the program will first allocate the total number of available sample points across the strata according to the Neyman allocation. Then, it will increase the allocation in each stratum to meet the minimum requirement, and remove the same number of points from the remaining strata according to selection rules that cause the least impact to overall precision.

Important note about fixed size sample designs: If you intend to supply a fixed number of sample points, you may find it useful to experiment with different numbers, to explore the relationship between

sample size and accuracy. For example, you may find that by adding only a few more sample points, accuracy can be improved significantly. Or, you may find that the fixed sample size you had in mind is actually larger than necessary to meet the desired accuracy. Results will vary with the population.

- **Number of strata**

  Within a single program run, you can request multiple analyses — each one using a different number of strata. This option applies to both fixed and optimal size sample designs.

About Using Prior Load Research Data (for Single Dimensional or Simple Random Designs)

> **Note:** If you are creating a multidimensional design, you will have the opportunity to apply prior load research data later in the design process (**Chapter 7**: **Recalculating Population Statistics for a Multidimensional Design Using Prior Load Research Data**).

The group mean and standard deviation within each stratum determine sample size. You have two options for calculating these statistics:

1. You can have the Sample Design Program calculate the mean and standard deviations for the stratification (usage) variable. However, since the stratification variable is only a proxy for the target demand, you cannot be sure how closely these statistics reflect actual variation in the target demand.
2. Alternatively, you can calculate the mean and standard deviations using actual measurement of the target demand, collected in prior load research studies. (You will use one of the Oracle Utilities Load Data Analysis programs to calculate the statistics.) Utilizing prior load research data, if it is available, can improve the quality of the design and possibly reduce the number of sample points.

When selecting existing load data for this purpose, be sure that the measured variable is closely correlated to your target variable. The data should be for a similar period, population type, and climate.

Using the Sample Design Program

Following is a summary of the steps you will use to stratify the population and determine sample size. Note that there are additional steps if you are incorporating load data from a prior study.

SUMMARY
STRATIFYING THE POPULATION AND DETERMINING SAMPLE SIZE FOR A SINGLE DIMENSION USING THE SINGLE DIMENSIONAL SAMPLE DESIGN PROGRAM (B310)

1. Verify or create the required input files:
   a. Control File — Frequency Distribution (.FDF) — use .FDF from B210, or create using another program.
   b. Sample Design Environment File (TGB31B.ENV) — parameters for your design.
3. Submit the job Single Dimensional Sample Design Program using the B310 Submit panel.
4. Check output.
   - Sample Design Environment Report (TGB310-01)
   - Frequency Distribution Table of Endpoint Block Usage (TGB310-02)
   - Sample Design Report(s) (TGB310-03).
5. Select sample design.
6. If you are incorporating prior load research data in your design, continue with the following additional steps (for single dimensional or simple random designs only).
7. Determine the mean and standard deviations for a prior sample using new strata breakpoints defined in Step 4. You will use either the Oracle Utilities Standard Load Analysis or Ratio Analysis Program to calculate these statistics.
8. Revise the Sample Design Environment File (TGB31B.ENV) with the mean and standard deviations from Step 5.
9. Resubmit the job (B310) with the new Environment File.
10. Check output and select a sample design.

Step 1A: Verify the Frequency Distribution File

Verify that the Frequency Distribution File (.FDF) has been properly set up and is located in your local DATA directory. If you have not already created the Frequency Distribution File (.FDF), see **Chapter 4**: **Analyzing the Population Frequency Distribution**.

Note: Do not specify zero as an endpoint of the distribution.

Step 1B: Create the Sample Design Environment File

You use the Sample Design Environment File to define how the population is to be stratified and sized.

As explained at the beginning of this chapter, there are many Environment File options available; you can use these options to create alternative sample designs. Within a single program run, you can specify alternative designs based on different numbers of strata (from 1 to 7). Other alternatives, such as fixed versus optimal sample size, require you to create different Environmental Files and re-run the program.

The file is made up of the comments shown in Figure 5-2. When creating the file, enter one command per line. Each line must begin with command name (keyword); but you need enter only the first three letters of the keyword (*except the MEAN Command, which requires all four letters*). Each keyword is followed by one or more user-specified parameters. You may enter the commands in any order. Command lines may be separated by blank lines; either commas or blanks may be used as delimiters between parameters. If you do not specify a command, the Sample Design Program will use the default value (underlined). Following is a detailed description of each command, and some examples of completed files.

> **Note:** Commands and parameters that appear in *italics* apply only when you are using statistics from prior load research, or when you are otherwise supplying your own strata breakpoints. These commands are explained in Step 6.

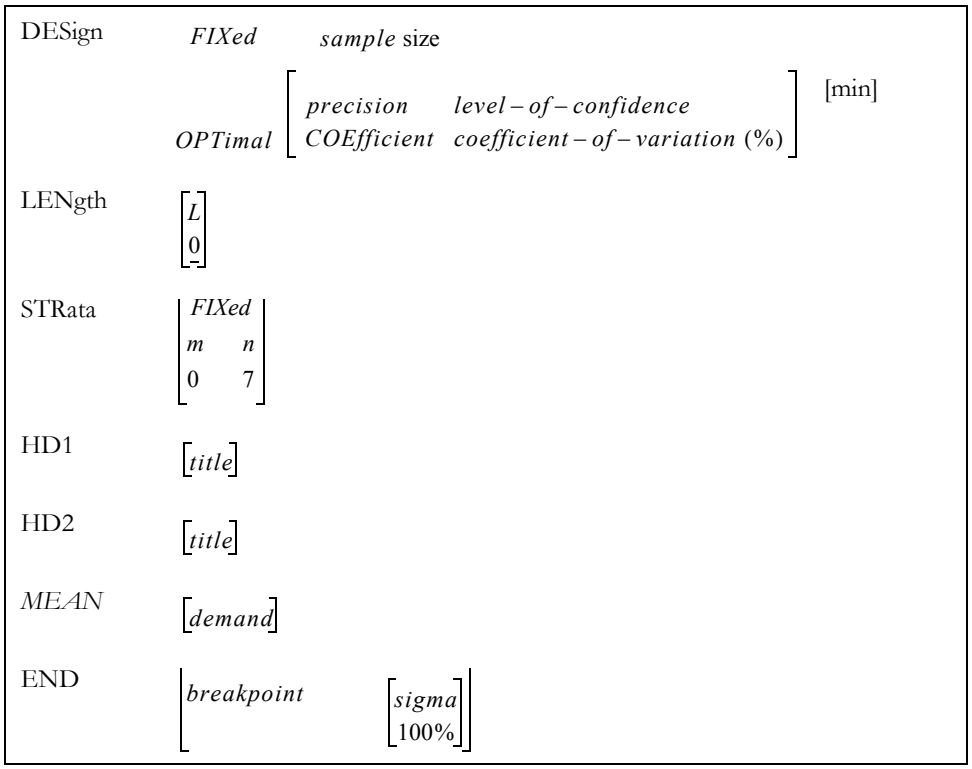


Figure 5-2 Sample Design Environment File Commands

- **Design** — A required command, which you use to establish the analysis as either a fixed sample size or an optimal sample size allocation. In a fixed size design, you supply the total number of desired sample points, and the program allocates the points to individual strata and calculates the resulting accuracy. In an optimal size design, you supply the desired accuracy and the program calculates the number of sample points necessary to obtain that accuracy.

If you wish to use a fixed sample size, enter **FIX** and the desired number of sample points (must be a positive number). For example, DES FIX 150.

If you wish to design an optimal sized sample, enter **OPT**, keyword with either of two formats. In the first format follow the OPT keyword with the desired precision expressed as an integer from 1 to 100 etc. and the desired level of confidence (either 90 or 95).

For example, if your sample requirement was a design accuracy of 10% at the 90% confidence level, you would create the following command: DES OPT 10 90. If you intend to validate your sample using the Validation Program, you are limited to either 5 or 10 for precision.

When you use this first format the coefficient of variance will be derived based upon the precision and the level of confidence specified.

The second format of the **OPT** command allows you to specify the coefficient of variance directly. This can allow better fine tuning of sample designs. Use the keyword **COE**, for example, DES OPT COE 2.55. The coefficient of variance must be between 1 and 100.

MIN parameter in the Design command — for either a fixed or optimal sized design, you can specify a minimum number of sample points to be allocated per stratum. For example, if you were creating a fixed sample with 150 points total and no less than 20 sample points per stratum, the Design command would be DES FIX 150 20. Or, if you wanted the optimal sample described in the previous paragraph with no less than 15 sample points per stratum, you would use the following Design command: DES OPT 10 90 15. (More information about how the program allocates the minimum number of points in a fixed or optimal sample is available. See Figure 5-1.)

- **Length** — Enter the lowest value for the stratification variable in the Population Data file. It must be a non-negative integer. The program will use this number as the starting point of the iterative calculations for the Dalenius-Hodges technique. You will typically use the default value, 0, since most populations will have some zero usage.
- **Strata** — Use this command to specify the number of strata to be analyzed in the program run. You must specify a lower (m) and upper (n) number for the range. For example, if you specify 1 and 3, the program will produce three sample designs, with 1, 2, and 3 strata, respectively.

  For each value in the range you supply, the program will perform a sample design (fixed or optimal depending on the design mode) to determine strata breakpoints and strata sample sizes. The first analysis is for the smallest number of strata (n). Each succeeding analysis is performed with an additional strata. The run terminates when the analyses exhaust the range of strata entered or a stratum sample size becomes zero. If you do not supply this command, the program will automatically generate a set of sample designs ranging from 1 to 7 strata.

  It is recommended that you specify the full range (all seven strata). Sample Design is an efficient program, so it is faster and easier to produce the widest range of alternatives in one run, rather than rerun the program if the first range you specify doesn't yield the desired results.
- **HDn** — The commands HD1 and HD2 allow you to label the sample design outputs. Both heading commands are optional.

  There must be a space between the command name (HD1 or HD2) and your title. The title can be any combination of printable characters. However, it can be no longer than 76 characters.
- **END** — Use this command to specify 100% sampling for selected, special strata. The format of this command is: END breakpoint 100%, where breakpoint is the value of upper boundary of the special stratum. For example, if you want to sample all customers in the stratum whose upper boundary is 123000, you would create the following command: END 123000 100%. (Note: In a fixed sample, the number of points required to satisfy 100% sampling for a specified strata will automatically be removed from the remaining strata according to selection rules that cause the least impact to overall precision.)

  The END Command can also be used with fixed strata boundaries. This application is described in Step 6.

Here are two examples of Sample Design Environment Files:

```
DES     OPT     10      95
LEN     0
STR     1       7
HD1     'SAMPLE DESIGN TEST CASE #4'
```

Figure 5-3 Sample Design Environment File for an Optimal Size Sample Analysis

```
DES     FIX     150     20
LEN     0
STR     1       3
HD1     'CERTIFICATION TEST CASE #2'
```

Figure 5-4 Sample Design Environment File

For a Fixed-Size Sample with a Minimum Requirement per Stratum (in this case, a total of 150 sample points with at least 20 points in each stratum).

Step 2: Submit the job

Use the B310 Submit panel.

Step 3: Check Output

The Sample Design Program produces three output files:

- **Sample Design Environment Report** — lists back the parameters you entered in the Environment File. Be sure to check this report closely to verify your input.
- **Frequency Distribution Table of Endpoint Block Usage** — presents a set of information on the stratification variable's frequency distribution, and the results of intermediate Dalenius-Hodges calculations for determining strategy breakpoints. (The formulas used by the program to arrive at these figures are listed in **Appendix B**: **Sampling Equations**.) You can use this report for verification of your work so far.
- **Sample Design Reports —** presents the results of the sample design. There will be one report for each number of strata in the range you specified in the Environment File Strata Command.

Step 4: Select Sample Design

Use the information provided in the Sample Design Reports to choose a sample design (number of strata and sample points).

In evaluating the reports, look for the design that yields the most desirable balance between accuracy and economy. You will usually look for the design that provides the required accuracy with the smallest number of points. (Usually, as the number of strata increases, the number of sample points decreases.)

Keep in mind that, depending upon the type of sample design you are creating, you will act on your selection in different ways (see Figure 5-1):

- *If you are incorporating load research data in your sample (single dimensional or simple random only)* — use the stratum breakpoints from the selected design to restratify an existing sample. After selecting a design, go to Step 5.
- *If you are creating a single dimensional stratified or sample random sample without incorporating load research data* — you will use your selection as the basis for choosing actual customers for the sample with the Sample Selection program. After you have selected a design, go to **Chapter 9**: **Selecting the Sample for a Single Dimensional Design**.
- *If you are creating a multidimensional stratified sample* — you will combine your selection with designs for the other dimensions. After you have run the Sample Design Program and selected a design (e.g., strata breakpoints) *for each usage variable in your design*, go to either **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics** if you intend to use the batch version of the software for the multidimensional process or

Appendix A: **The User Language** if you intend to use the Interactive Multidimensional Sampling Package.

Step 5: Determine the Mean and Standard Deviations

You need to determine the mean and standard deviations for a prior sample using new strata breakpoints. Utilizing statistics based on an existing load research study can improve your new sample design. In this step, you will use other Oracle Utilities Analysis programs to restratify an old sample according to your new strata breakpoints, and compute the group mean and strata standard deviations at the design hour for input into Sample Design. Here is what you do:

- Note the strata breakpoints for your selected design in the Sample Design Report (these values are labeled "Endpoint of Stratum Range").
- Be sure that the sample data is available in the ELDB — Oracle Utilities Load Database. See *Chapter 5* in the *Oracle Utilities Load Analysis Load Data Analysis User's Guide* for more information.
- Run either the Oracle Utilities Load Data Analysis Program or the Ratio Analysis Program on the existing load data using the new breakpoints. Use the same program that will ultimately be used to analyze the data collected for your new sample. See *Chapter 7* in the *Oracle Utilities Load Analysis Load Data Analysis User's Guide* for additional information about setting up and running these programs.

  Here are some important points to keep in mind when creating your Analysis Environment File:

  Enter the new breakpoints in the STRATUM Command.

  Use the ASSIGN FLOAT Command so old stratum assignments are ignored.

  Specify the design hour or hours of interest (typically the hour of System Peak) with the PEAK Command.

  Specify WRITE NO, because you need only a paper copy of the statistics (typically, you won't want to save the statistics in the Oracle Utilities database).

- At the end of the analysis run, refer to the following reports for the desired statistics. Remember, you're looking for the sample (group) mean and the standard deviation for each stratum. E.g., if you had three strata, you would look for a total of four values.

  If you used Standard Load Analysis:

  TGY311-10 "Maximum Coincident Demand Report"

  For the entire analysis population, note the value for "Sample Mean" at the time of the supplied peak.

  TGY311-80 "Standard Deviation of Sample Demand"

  For each stratum, note the value for "S.D. of Demand" at the time of the supplied peak.

  If you used Ratio Analysis:

  TGY331-10 "Maximum Coincident Demand Report"

  For the entire analysis population, note the value for "Sample Mean" at the time of the supplied peak.

  TGY331-80 "Standard Deviation of Sample Residuals of Demand"

  For each stratum, note the value for "S.D. of Residuals" at the time of the supplied peak.

Step 6: Revise the Sample Design Environment File

Revise the sample design environment file with the mean and strata standard deviations from step 5. You will now revise the file you created in Step 1B by adding or updating the following three commands. (***Note:*** You can, of course, change the titles using the HD1 and HD2 commands; however, do not change the Design or Length commands.)

- **Strata** — For defining strata boundaries the input should be: **STR FIX.**
- **Mean** — Input the value for the sample mean.
- **End** — For each stratum in your design, input the stratum breakpoint and the standard deviation. You **must** supply one END Command for each stratum, and the breakpoint for the last END Command **must** be "INFINITY".

Here is an example of a revised Sample Design Environment File:

```
DES     OPT     10      95
LEN     0
STR     FIX
HD1     'SAMPLE DESIGN TEST CASE #5'
MEAN    5.5646
END     2200.00 1.6967
END     4900.00 1.1717
END     15000.00 2.8561
END     INFINITY 1.567
```

Figure 5-5 Sample Design Environment File

Revised to include the Group Mean and Stratum Standard Deviations from a prior Sample.

Step 7: Resubmit the Job with the New Environment File

Use the B310 Submit panel.

Step 8: Check Output and Select a Sample Design

Evaluate your output and select a design using the same instructions and criteria provided in steps 3 and 4.

After you have chosen a design, you will be ready to draw a list of actual customers for the sample using the Sample Selection Program. See **Chapter 9**: **Selecting the Sample for a Single Dimensional Design**.

Chapter 6

Assigning the Population to Cells and Calculating Population Statistics

After you have used the Single Dimensional Sample Design Program to determine the strata breakpoints for each usage variable in your multidimensional design, you are ready to bring all of the design dimensions together to create cells (the intersections of the strata) and to determine in which cell each customer belongs. You will use the **Multidimensional Population Analysis Program** for this process.

When you run Population Analysis, the program also calculates the number of customers in each cell and the cell means and standard deviations for one stratification (usage) variable. The program writes these statistics to the Population Statistics File, which is a required input for the next step, Multidimensional Sample Design and a later step, Sample Validation.

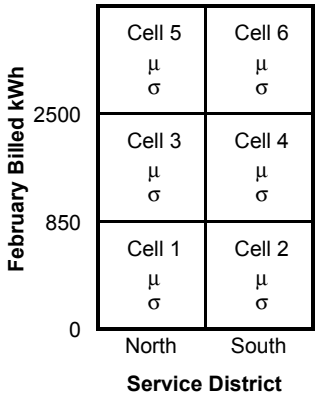
Keep in mind that the Multidimensional Sample Design and Sample Validation Programs require a different set of cell statistics for each usage variable in your design. If you intend to use the cell statistics calculated by the Population Analysis Program as the basis for sizing and/or validating the sample, you must run the Population Analysis Program once for each usage variable in your design. (See Figure 6-1).

Sizing a Multidimensional Sample Using Prior Load Research

Statistics calculated by the Multidimensional Population Analysis Program are typically based on a proxy variable. Alternatively, you can calculate the cell means and standard deviations required for sizing the sample by applying one of the Oracle Utilities Load Data Analysis programs to prior survey data for the actual target demand. (This process can improve the quality of your design and is described in **Chapter 7**: **Recalculating Population Statistics for a Multidimensional Design Using Prior Load Research Data**.) Even if you plan to use that approach, you still must first run the Multidimensional Population Analysis Program as described in this chapter to assign customers to the appropriate cells, develop cell population counts, and generate population statistics for use in the Sample Validation phase.

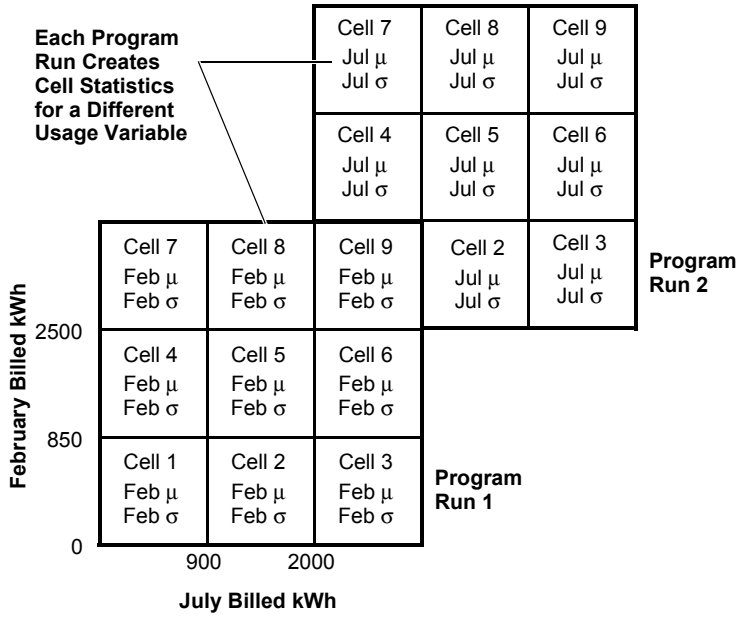
Example 1

Two-Dimensional Design with One Usage Variable and One Demographic Variable



Example 2

Two-Dimensional Design with Two Usage Variables



The Multidimensional Population Analysis Program assigns each member of the population to the appropriate cell and calculates cell statistics for one usage variable at a time. If your design has one usage variable, you will run the program just once. If your design has two or more usage variables, you must run the program two or more times, to calculate the cell means and standard deviations for each usage variable.

Figure 6-1 Assigning the Population to Cells and Calculating Cell Statistics

Following is a summary of the steps you will follow to assign customers to the correct cells and create the Population Statistics File(s).

SUMMARY
ASSIGNING THE POPULATION TO CELLS AND CALCULATING POPULATION STATISTICS USING THE MULTIDIMENSIONAL POPULATION ANALYSIS PROGRAM (B220)

1. Verify or create the required input files:
   - Population Data File (.PDF), located in the COMMON\DATA directory on the server.
   - Record Definition File (.DEF) — the output file created by B110.
   - Population Analysis Control File (TGB22A.CTL) — logic required to identify and assign population to cells and calculate statistics, based on information from Single Dimensional Sample Design Report.
2. Submit the job (Multidimensional Population Analysis Program — B220).
3. Check output:
   - Population Analysis Execution Log (TGB220-01)
   - Table of Cell Entries (TGB221) — report of population statistics.
   - New copy of the Population Data File updated with STRATA numbers (TGB222) — will be used as Population Data File for B420, Multidimensional Sample Selection in **Chapter 10**: **Selecting the Sample for a Multidimensional Design**. Name this file differently from the original .PDF, so you don't overwrite it when you copy it to your COMMON\DATA directory.
   - Population Statistics File (.PSF) — same data as TGB221, but formatted for input to B320, Multidimensional Sample Design in **Chapter 8**: **Determining Sample Size for a Multidimensional Sample Design** and B520, Sample Validation in **Chapter 11**: **Validating the Sample** (TGB32A and TGB52C, respectively).

If your design has more than one usage variable:

4. Repeat the steps listed above to calculate cell statistics for each additional usage variable in your design.

Important note to those with more than one usage variable: As explained earlier, the Population Statistics File output by the Population Analysis Program is a required input to the next step, Multidimensional Sample Design, and a later step, Sample Validation. You must create one Population Statistics File for each usage variable in your design. However, when you rerun Population Analysis for the next variable in your design, the program will write over the existing file. **Be sure to preserve existing files by assigning a new data set name to the Population Statistics File each time you run the Population Analysis Program for a different usage variable**.

Warning about space requirements! Each time you run Multidimensional Population Analysis, the program will create another copy of the Population Data File, this one with cell assignment numbers placed in the STRATA field. The file can be very large. Have your Oracle Utilities Load Analysis Administrator make sure you have enough disk space.

Step 1A: Verify that the Population Data File is Available

Verify that the Population Data File (.PDF) has been properly set up and is available. It should be available for selection in the submit panel under the Population Data file dropdown. If you have not already created the Population Data File (.PDF), see **Chapter 3**: **Creating the Population Data File and Record Definition File**.

Step 1B: Verify that the Customer Record Definition File exists

If you have not already created the Record Definition File (.DEF), see **Chapter 3**: **Creating the Population Data File and Record Definition File**.

Step 1C: Create the Population Analysis Control File

The objectives of this file are to specify the number of dimensions, strata, and special cells in your multidimensional design, and to provide the logic that the program needs to identify the population and classify it into the appropriate cells (including strata breakpoints for usage variables and/or category definitions for demographic variables). The Population Analysis Control File commands are as follows:

SELect	field = value
USAge	usageVariable
DIMension	dimension1, dimension2
BREakpoints	dim1Boundary1, dim2Boundary1; … dim1BoundaryN, dim2BoundaryN; INF, INF

SELect	field = value

SELECT (optional) - SELECT is a test statement for defining eligible customers. For example, to process records whose RATE code is 'SGS' only: SELECT RATE = 'SGS'

USAge	usageVariable

USAGE - Defines the usage, or design variable. Specify the name of the usage variable you would like to use. For example:

USAGE JAN

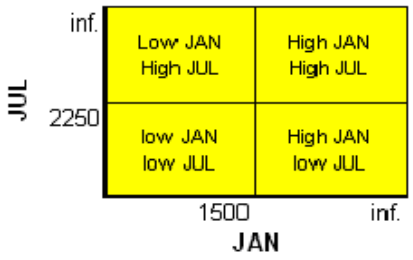
DIMension	dimension1, dimension2

DIMENSION – Specify the usage or design variable for each dimension in your design. If you have two dimensions then you will have two values here.

BREakpoints	dim1Boundary1, dim2Boundary1; … dim1BoundaryN, dim2BoundaryN; INF, INF

BREAKPOINTS – Define the lower stratum breakpoints for each cell. Each breakpoint is comma separated list of lower boundaries for each dimension and each breakpoint is semi-colon separated. Special Cells are enclosed in parenthesis "()". The last breakpoint must be INF. (infinity) You should already have the breakpoints determined from your prior step.

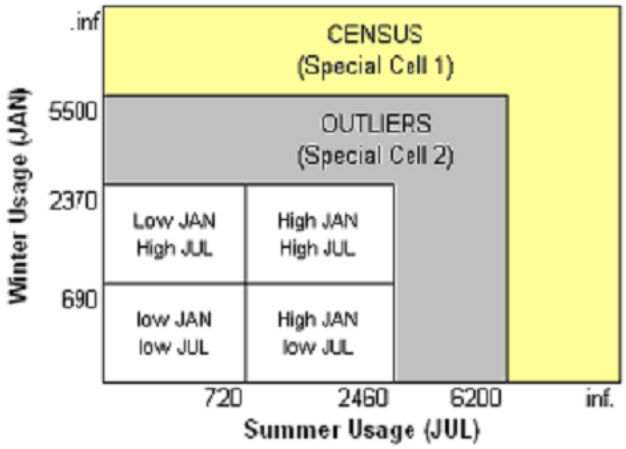
Example: To create the following Multidimensional design (with no special cells):



We would construct the following control file:

```
USAGE   JUL
DIMENSION    JAN, JUL
BREAKPOINTS     1500, 2250; INF, INF
```

Example: To create the following Multidimensional design with Special Cells:



We would construct the following control file:

```
USAGE   JUL
DIMENSION    JUL, JAN
BREAKPOINTS      720,690;2460, 2370;(6200, 5500);(INF, INF)
```

Create the Population Analysis Control File with User Language Statements (Optional)

You can optionally compose the Population Analysis Control File with User Language statements. If you already have constructed your analysis control file, you can skip this section. The "User Language" allows for more customization, but keep in mind it is also more complex.

The objectives of this file are to specify the number of dimensions, strata, and special cells in your multidimensional design, and to provide the logic that the program needs to identify the population and classify it into the appropriate cells (including strata breakpoints for usage variables and/or category definitions for demographic variables).

Important! Remember, you must create a different Control File and run the Population Analysis Program for each usage variable in your multidimensional design. *However, the only modification you must make to the Control File between program runs is to identify a different usage variable as the design variable in the DIM Statement (referred to as "value" in the explanation of DIM statements provided below).*

Here are example Control files for your review (Figure 6-2 through Figure 6-5). You may find it easiest to create your own file by simply modifying one of the examples according to the specifics of your project.

As you can see, many of the elements are the same or similar to those you have probably already used to create other sampling Control Files, such as the Analysis Control File in **Chapter 4**: **Analyzing the Population Frequency Distribution**. The following explains the elements of the User Language which are specific to the Population Analysis Control File.

- **DIM Statement** — You must include one DIM Statement in your Population Analysis Control File. This statement specifies the total number of dimensions in your multidimensional design, the number of strata within each dimension, the number of special cells, and the usage variable for which the program will calculate the cell means and standard deviations in the program run. Use this format to construct it:

  DIMn a_1 a_2...a_n b value

where:

n is the number of dimensions in your design.The value must be an integer between 1 and 7.

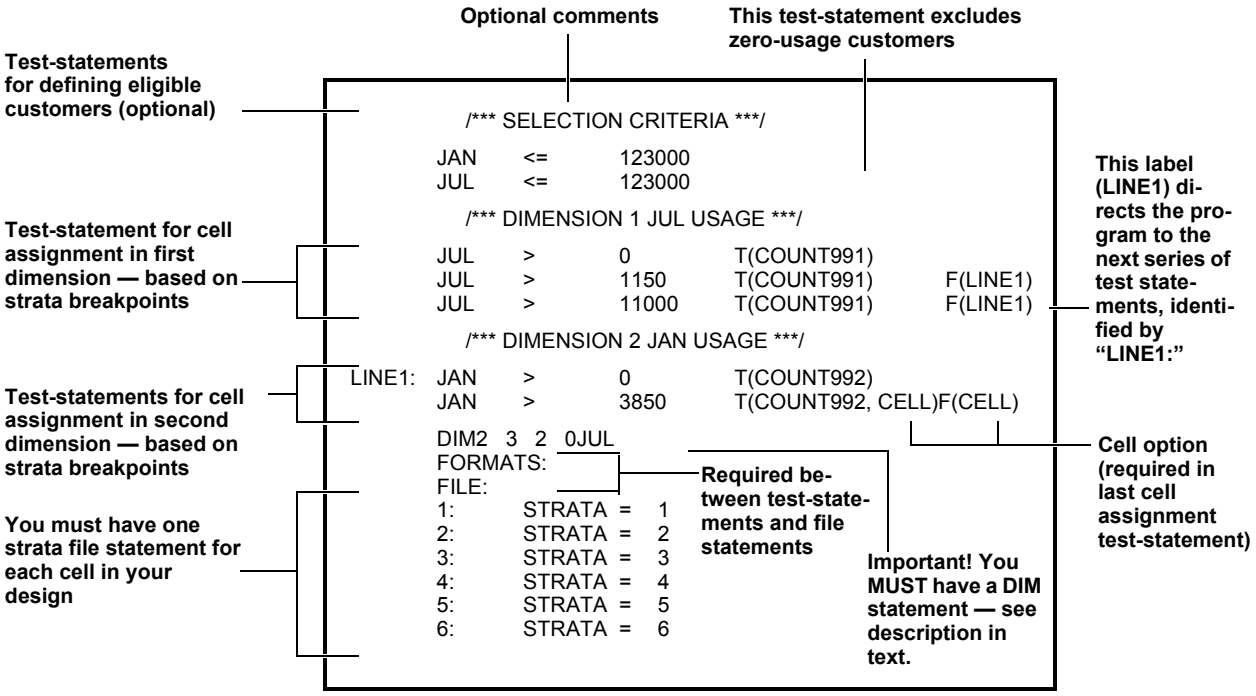


Figure 6-2 Example of a Population Analysis Control File

For a Multidimensional Design with Two Usage Variables (January and July Billed Energy)

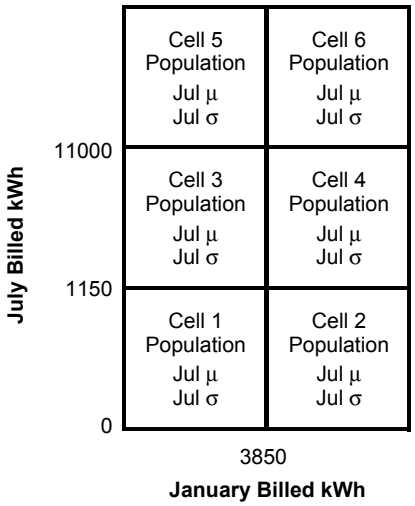


Figure 6-3 Visual Representation of Cell Assignment Scheme and Cell Statistics

Created with the Control File in Figure 6-2.

Note: Cell statistics will be created for July, because JUL is indicated in the DIM Statement.

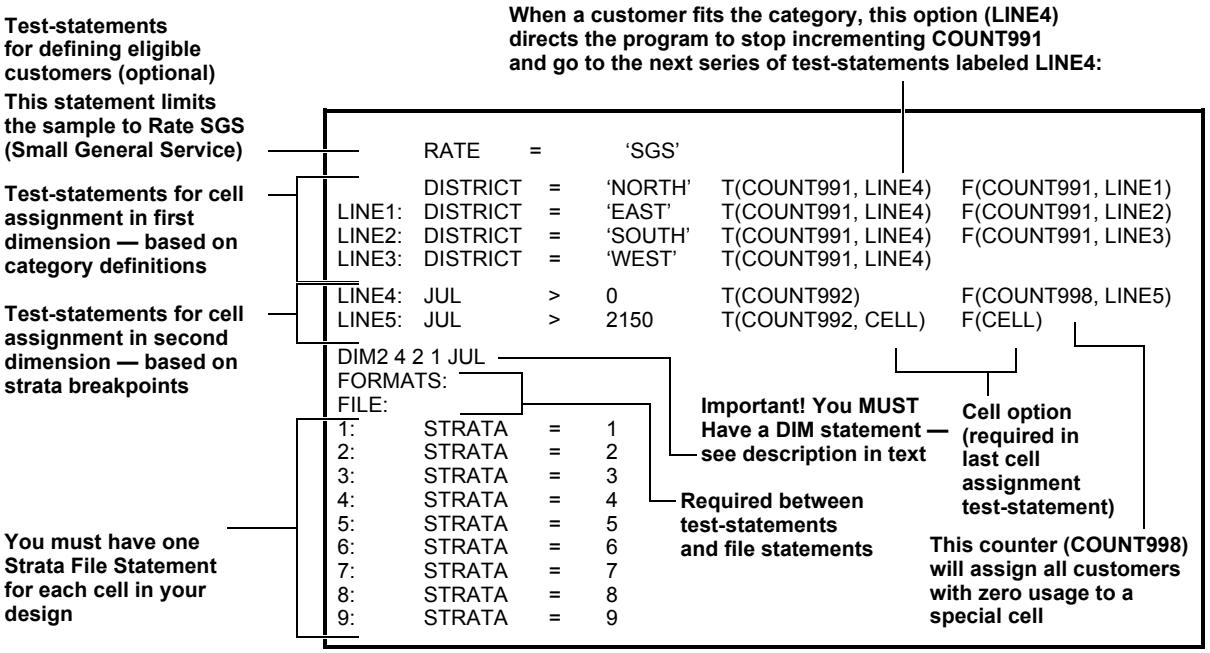


Figure 6-4 Example of a Population Analysis Control File

For a Multidimensional Design with One Demographic Variable (DISTRICT), One Usage Variable (JUL), and a Special Cell (for Customers with Zero Bills).

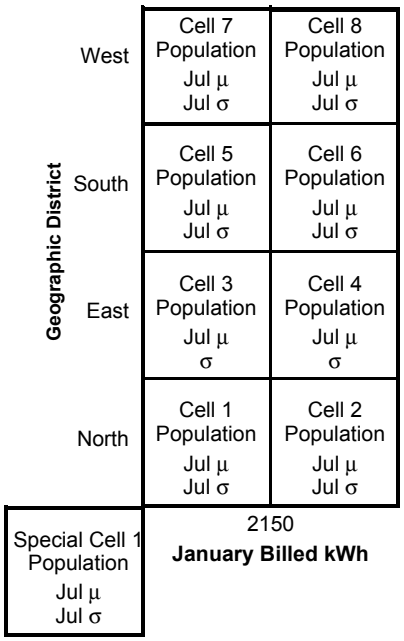


Figure 6-5 Visual Representation of Cell Assignment Scheme and Cell Statistics

Created with the Control File Shown in Figure 6-4.

a_1 — a_n is the number of strata in each dimension. There must be n values and each an integer between 2 and 9, inclusive.

b is the number of special cells in your design. You can specify up to 2 special cells. A value of zero must be used if there are no special cells.

value is the usage variable in the Population Data File for which the statistics (population count, mean, and standard deviation for each cell) are to be computed in the program run. Be sure to spell the identifier for the usage variable the same way it was spelled in your Record Definition File.

Important! Do *not* place a blank between DIM and n. Do place one or more blanks between all other parameters in the statement. Correct example:

DIM2 3 2 1 JUL

Note: The total number of cells (computed as $a_1 * a_2 ... * a_n + b$) must be no greater than 200.

For example, let's say that you wish to construct a DIM statement for a 2 dimensional design. The first dimension in your design contains 3 strata, and the second dimension contains 2 strata; and you have designated one special cell for customers with very high usage. "JUL" is the name of the usage variable in the Population Data File for which you wish to calculate cell statistics in this program run. You would construct the DIM statement as follows:

DIM2 3 2 1 JUL

- **Counters 991 through 997** — use counters 991-997 to construct the logic for cell assignment. Use one counter per dimension, e.g., COUNT991 for dimension 1, COUNT992 for dimension 2, etc.

  Essentially, you set up a series of test-statements for each dimension. Counters — special options within the test-statements — keep track of the number of times a customer passes or fails a test-statement. The program will assign each customer to a cell according to the number of counts it receives.

  How you structure the test-statements and counters depends upon whether the dimension is a usage or demographic variable.

  For dimensions based on usage variables — each test statement identifies the lower boundary of a stratum in the dimension. (Use the strata breakpoints you determined with the Single Dimensional Sample Design Program. Create each test statement with this format:

  usage variable > strata lower bound T(COUNT99n) F(label)

  where n is the number of the dimension; e.g., 1 for the first dimension, 2 for the second dimension, etc.

  Which means, if the customer's value for the usage variable is greater than the value for the strata lower bound, increment counter 99n by 1. If it is less than the strata breakpoint, go to the next series of test-statements identified by a label, such as LINE1. (See Figure 6-1.)

  During the program run, each counter counts the number of times a customer passes a test-statement until it finally fails and falls into a "bucket," e.g. the stratum in which it belongs for that variable (e.g., dimension).

  Note: Typically, the first test-statement in a series should be:

  usage variable > 0 T(COUNT99n).

This excludes zero-usage customers from your sample. (Refer to Figure 6-4 to see how to put zero-usage customers in a special cell.)

For dimensions based on demographic or categorical variables (such as SIC code, service territory, etc.) — each test statement identifies one category or group. Use the format:

demographic variable = 'value' T(COUNT99n, label for next series) F(COUNT99n, label for next statement)

Which means, if the customer's value for the demographic variable matches the specified category, increment counter 99n by one and stop testing for that dimension (the customer is in its appropriate "bucket"). If not, increment counter 99n by one and go to the next test-statement in the series. (Figure 6-2)

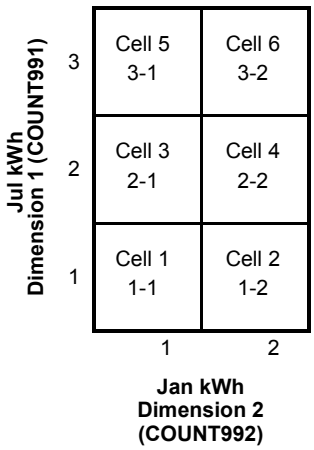
Important note: If the variable was defined as character type data in your Record Definition File, you must enclose the value in your test-statement in single quote marks. For example, DISTRICT = 'NORTH' or RATE = 'SGS'.

Remember, all test-statements in a series use one counter — that is, all test-statements for dimension 1 use COUNT991, all test-statements for dimension 2 use COUNT992, etc.

During the program run, the program combines the counts it gets for all counters (e.g., dimensions) to arrive at a cell assignment. This process for assigning three customers to cells in a two-dimensional design is visually represented below:

Two-Dimensional Example (3 x 2)

	JUL kWh COUNT991	JAN kWh COUNT992	Cell Assignment
Customer #1	1	2	2
Customer #2	3	1	5
Customer #3	3	2	6



- **Counter 998 and 999 (optional)** — Use COUNT998 and COUNT999 if your design includes special cells. You can have up to two special cells. Use COUNT998 for the first special cell, and COUNT999 for the second. (See Figure 6-4.)

  If COUNT998 or COUNT999 are triggered by a test-statement, then the program ignores all other counters and assigns the customer to the appropriate special cell.

- **False clause** — In most test-statements, you must include a False clause that directs the program to the next series of test-statements when a customer fails the statement. In those cases, construct the false clause with a "label" that directs the program to the first line of the next series of test-statements. For example, in Figure 6-2, the false clauses in the first set of cell assignment test-statements point the program to the statement labelled "LINE1".

  There are cases when you would not include a false clause with a label in a test-statement. Since the default False clause is NEXT, this is an easy way to exclude customers that fail a given test statement. For example, if you wished to exclude customers with zero-usage for July, your first test-statement might be: JUL > 0 T(COUNT991)

- **Cell option** — You must include the CELL option in your last test-statement, and it must be specified as both the True and False action. This option says to the program, the logic for assigning population members to the cells is complete, now execute it.

- **Strata File statements** — The Strata File statements create a new Population Data Record for each customer, with the customer's cell assignment number placed in the STRATA field.

  You must construct one Strata File statement for each cell in your design, including special cells. (Although the term STRATA must be used, the statements actually refer to cells.) Use the following format:

  ```
  FORMATS:
  FILE:
  1: STRATA = 1
  •
  •
  •
  N: STRATA = N
  ```

  where N is the total number of cells in your design, computed as $a_1 * a_2 ... * a_n + b$ (see description in DIM statement earlier).

  Be sure to precede the Strata File statements with the following two keywords, one per line:

  ```
  FORMATS:
  FILE:
  ```

Step 2: Run the Multidimensional Population Analysis Program

Use the B220 Submit screen.

Step 3: Check Output

The Multidimensional Population Analysis Program produces four outputs:

- **Population Analysis Execution Log (TGB220-01)** — A printed version of your Population Analysis Control File. If the program encounters any errors during processing, they will be listed in the "Extracted Data Report" portion of the log. Be sure to check the Execution Log carefully.
- **Table of Cell Entries (TGB221)** — reports the number of customers assigned to each cell, along with the mean and standard deviation for the design (usage) variable in each cell. The report flags any empty cells or cells with only one customer. Also, empty cells will have the mean and standard deviation set to zero, and cells with only one customer will have the standard deviation set to zero.

Report version of Population Statistics File (format differs slightly between report and file versions).

- **Scored Population Data File (.PDS)** — an updated version of the Population Data File. The STRATA field for each member of the population now includes a cell assignment number. This file is used as input to the Multidimensional Sample Selection Program (B420).
- **Population Statistics File (.PSF)** — same data as TGB221, but formatted for input to Multidimensional Sample Design (B320) and Sample Validation (B520). After a successful run, this file should be available for selection in Multidimensional Sample Design (B320) and Sample Validation (B520).

Step 4: Repeat the Steps Listed Above

Repeat the steps listed above to calculate population cell statistics for each additional Usage Variable in your design. Population statistics for each usage variable are a required input to Multidimensional Sample Design and Sample Validation.

Therefore, you must run the Population Analysis Program once for each usage variable in your design. As explained earlier, only two modifications are required between program runs:

- You must identify a different usage variable in the Control File's USAGE statement. (or DIM statement if you are using the User Language).
- You must assign new filenames to the Population Statistics Files (.PSF) to avoid overwriting statistics calculated in earlier runs.

Chapter 7

Recalculating Population Statistics for a Multidimensional Design Using Prior Load Research Data

In the previous chapter, you applied the Multidimensional Population Analysis Program to assign customers to the appropriate cells and calculate cell population statistics. You can use these statistics as the basis for sizing your sample with the Multidimensional Sample Design Program in **Chapter 8**: **Determining Sample Size for a Multidimensional Sample Design**. However, the Population Analysis Program calculates cell statistics using a proxy variable that only approximates the target demand. To improve the quality of your design and possibly reduce the number of sample points, we recommend that you recalculate the cell means and standard deviations using existing actual load research data for a similar population. If you do not have prior survey data, you may be able to borrow some from another utility with a similar customer base and climate.

Essentially, you will post-stratify the old sample according to your new cell definitions; then compute new cell statistics for the design hour(s) using one of the **Oracle Utilities Load Data Analysis programs**.

SUMMARY
RECALCULATING POPULATION STATISTICS FOR A MULTIDIMENSIONAL DESIGN USING PRIOR LOAD RESEARCH DATA AND ORACLE UTILITIES LOAD ANALYSIS PROGRAMS (Y310 OR Y330)

1. Manually post-stratify the sampled population according to new cell definitions.
2. Run one of the Load Analysis Programs (Y310 or Y330) to calculate new cell means and standard deviations — must be performed for each usage variable in the design.
3. Update the Population Statistics Files (.PSF from B220) with the new statistics — one file per usage variable in your design.

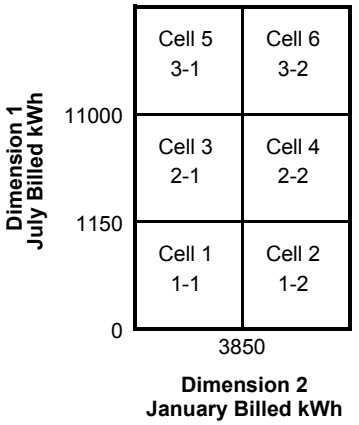
   Note: Be sure to save a copy of the original version of the Population Statistics File(s), because you will need them for Validation.

Step 1: Manually Post-Stratify the Sampled Population

You need to manually post-stratify the sampled population according to new cell definitions. In this step, you will manually assign a cell number for each member of the sampled population. You will later use these numbers to construct your Oracle Utilities Analysis Control File — in other words, to tell the Analysis programs in which cell each customer belongs.

First, look up each customer's value for every dimension, both usage and demographic. For dimensions based on usage variables, look up the same proxy variable that you originally used to stratify the target population, typically billed energy.

Based on that information, assign the appropriate cell number to each customer. Use the same cell numbering scheme that the Multidimensional Population Analysis Program applied to the target population in the Population Statistics File (based on your Population Analysis Control File). You may find it helpful to visualize the numbering scheme by drawing a grid that represents the intersection of the strata and/or demographic segments in each dimension. Number the resulting cells from 1 to n, where n represents the total number of cells in your design. An example for a two-dimensional design is shown in Figure 7-1 (this example corresponds to the Population Analysis Control File in Figure 6-2).



Assign Customer B007:
July kWh = 1300 kWh Strata 2 in First Dimension
January kWh = 3000 kWh Strata 1 in Second Dimension
2-1 = Cell Number 3

Figure 7-1 *Draw a Grid to Represent the Cells in Your Multidimensional Design and Assign Each Customer a Cell Number*

Step 2: Run One of the Load Analysis Programs

You must run one of the load analysis programs to calculate new cell means and standard deviations. In this step, you will calculate the new cell means and standard deviations for the sample data at the design hour(s) using one of the Oracle Utilities Load Analysis Programs, or another program of your choice. The following instructions assume you are using a Oracle Utilities Load Analysis program.

First, identify which analysis program you intend to use to calculate the new statistics. Two Oracle Utilities Load Analysis programs are available — **Standard Load Analysis** for mean per unit calculations and **Ratio Analysis** for ratio expansion. *Use the same program that will ultimately be used to analyze the data collected for your new sample.*

See the Oracle Utilities Load Analysis Load Data Analysis User's Guide for detailed instructions on how to set up and run these two programs. Here are some important points to keep in mind when creating your inputs:

- Be sure that the sample data is available in the ELDB — Oracle Utilities Extracted Load Database (see *Chapter 5* in the *Oracle Utilities Load Analysis Load Data Analysis User's Guide* for more information).
- When creating the Analysis Control File, use the cell numbers you developed in Step 1 for "stratum-number."

- When creating the Analysis Environment File:

  Use the ASSIGN FIXED Command, because you want the customers assigned to the cells you identified in Step 1.

  Specify the design hour or hours of interest (typically the hour of System Peak) with the PEAK Command.

  Specify WRITE NO, since you need only a paper copy of the statistics (typically, you do not want to save the statistics in the Oracle Utilities ELDB.)

  At the end of the analysis run, refer to the following reports for the desired statistics. Remember, you're looking for the means and standard deviations for each cell. ***Note:*** In these reports, "Strata" refers to "Cell."

  If you used Standard Load Analysis:

  TGY311-10 "Maximum Coincident Demand Report"

  For each stratum (e.g., cell), note the value for "STRATA MEAN" at the time of supplied peak.

  TGY311-80 "Standard Deviation of Sample Demand"

  For each stratum (e.g., cell), note the value for "S.D. of Demand" at the time of supplied peak.

  If you used Ratio Analysis:

  TGY331-10 "Maximum Coincident Demand Report"

  For each stratum (e.g., cell), note the value for "STRATA MEAN" at the time of supplied peak.

  TGY331-80 "Standard Deviation of Sample Demand"

  For each stratum (e.g., cell), note the value for "S.D. of Demand" at the time of supplied peak.

Step 3: Update the Population Statistics File (.PSF)

Return now to the Sampling Package. With the new cell means and standard deviations you just developed, update copies of the Population Statistics File (.PSF) you originally created in **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**[1]. If you have more than one usage variable in your design, make additional copies of the file and update each with the new statistics. There must be a different file for each usage variable in your design.

Alternatively, if you do not wish to work with copies of the Population Statistics File, you can create the new file(s) from scratch using the following format. Each line of the file must contain the data for one cell. The entries must be in ascending order by cell number. All cells in the analysis must be included, even if the cell is empty.

cell-id, cell-number, cell-population, cell-mean, cell-standard-deviation,

- **cell id** — combination of the stratum numbers in each dimension that defines the cell; should be the same as that shown on the original Population Statistics File produced by B220.

1. **Important Note:** If you plan to validate your sample, be sure to save the original versions of the Population Statistics Files you created with the Multidimensional Population Analysis Program in **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**. Validation should be based on the target population.

- **cell number** — use same numbering scheme you used to post-stratify the sampled population; should also be the same as that shown on the original Population Statistics File produced by B220.
- **cell population** — a count of the number of customers in the Population Data File which were assigned to that particular cell. This information can be obtained from the original Population Statistics File or "Table of Cell Entries" report produced by B220.
- **cell mean** — statistic just computed for the design variable.
- **cell standard deviation** — statistic just computed for the design variable.

An example Population Statistics File is shown in Figure 7-2.

```
CELL(1-2),    1,     173,    484.21,     406.52,
CELL(1-2),    2,       2,   3699.00,    2018.08,
CELL(2-1),    3,     672,    417.15,     458.02,
CELL(2-2),    4,      52,   4949.02,    4497.65,
```

Figure 7-2 Example of a Population Statistics File

Chapter 8

Determining Sample Size for a Multidimensional Sample Design

In this step, you will determine the sample size for your multidimensional design. You will first apply the **Multidimensional Sample Design Program** to determine the number of sample points required to meet the desired accuracy for each usage variable in the design; you will then manually combine the cell sizes for all usage variables to define a sample that satisfies accuracy requirements for all dimensions simultaneously.

The Multidimensional Sample Design Program reads the Population Statistics File (.PSF), calculates the group mean from the supplied sample means, and computes the optimal sample size based on the sum of the weighted standard deviations and the required level of precision. The Neyman allocation procedure is used to determine the sample size for each cell.

You have the option of allocating a fixed sample size, or of supplying a required precision and level of significance to determine an optimal sample size. Other design options: you can specify a minimum number of points per cell; and you specify that all points in a given cell be included in the sample (100% sampling).

You must run the Sample Design Program for each usage variable in your design. The only difference between the program runs is the Population Statistics File you use as input.

Following is a summary of the steps you will follow.

SUMMARY
DETERMINING SAMPLE SIZE FOR A MULTIDIMENSIONAL SAMPLE DESIGN USING THE MULTIDIMENSIONAL SAMPLE DESIGN PROGRAM (B320)

1. Verify or create the required input files:
   a. Control File — Population Statistics File (.PSF). Use the file from B220, possibly modified with statistics for prior load research data. You may also wish to specify 100% cell sampling.
   b. Sample Design Environment File (TGB32B.ENV) — parameters of your design.
3. Submit the job Multidimensional Sample Design Program — (B320).
4. Check output:
   - Multidimensional Sample Design Environment Report (TGB320-01)
   - Multidimensional Sample Design Execution Log (TGB320-02)
   - Multidimensional Sample Design Report (TGB320-03)
   - Multidimensional Sample Design Summary Report (TGB320-04)
   - Relative Accuracy File (.RAF) — will be used as Environment File (TGB52B.RAF) for B520, Sample Validation in **Chapter 11**: **Validating the Sample**.

If your design has more than one usage variable:

5. Repeat the preceding steps for each additional usage variable in your design (using a different Population Statistics File).
6. Manually combine cells.

Step 1A: Verify the Population Statistics File

You need to verify that the population statistics file has been properly set up and is located in your DATA directory (specify 100% sampling, if desired). The Population Statistics File is initially created with the Multidimensional Population Analysis program (.PSF from B220). If you have not already created this file, refer to **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**. Alternatively, you can use a version of this file that has been modified to reflect statistics computed from prior survey data (see **Chapter 7**: **Recalculating Population Statistics for a Multidimensional Design Using Prior Load Research Data**).

Important Note about 100% Sampling: If you wish to include all members of a given cell in your sample, you must specify the cell or cells in the Population Statistics File. Type 100% in columns 77 — 80 of the row for the cell. The % sign is required. In the example below (Figure 8-1), all members of Cell 4 will be sampled.

```
CELL (1-1),    1,    173,    484.21,    406.52,
CELL (1-2),    2,      2,   3699.00,   2018.08,  100%
CELL (2-1),    3,    672,    417.15,    458.02,
CELL (2-2),    4,     52,   4949.02,   4497.65,  100%
```

Figure 8-1 Example Population Statistics File with 100% Cell Sampling Specified

Step 1B: Create the Sample Design Environment File (TGB32B.ENV)

You use the Sample Design Environment File to specify either a fixed or optimal size sample, and to assign titles for the Multidimensional Sample Design outputs. Following is a detailed description of each command used to create this file, and some examples of completed files.

DESign {**FIXed** *sample-size* | **OPTimal** *precision level-of-confidence* | **OPTimal COEfficient** *coefficient-of-variation*} [*min*]

HD1 [*title*]

HD2 [*title*]

- **Design** — A required command, which you use to establish the analysis as either a fixed sample size or an optimal sample size allocation. *If you wish to use a fixed sample size*, enter **FIX** and the desired number of sample points (must be a positive number). For example, DES FIX 150.

  If you wish to design an optimal sized sample use the **OPT** keyword with either of two formats. In the first, follow the **OPT** keyword with the desired precision expressed as an integer from 1 to 100 etc. and the desired level of confidence (either 90 or 95). For example, if your sample requirement was a design accuracy of $\pm$ 10% at the 90% confidence level, you would create the following command: DES OPT 10 90. Note: If you intend to validate your sample using the Validation Program, you are limited to either 5 or 10 for precision.

  When you use this first format the coefficient of variance will be derived based upon the precision and the level of confidence specified.

  The second format of the **OPT** Command allows you to specify the coefficient of variance directly. This can allow better fine tuning of sample designs. Use the keyword **COE**; for example, DES OPT COE 2.55. The coefficient of variance must be between 1 and 100.

  min parameter in the Design Command — for either a fixed or optimal sized design, you can specify a minimum number of sample points to be allocated per cell. For example, if you were creating a fixed sample with 150 points total and no less than 20 sample points per cell, the Design Command would be DES FIX 150 20. Or, if you wanted the optimal sample described in the previous paragraph with no less than 15 sample points per cell, you would use the following Design Command: DES OPT 10 90 15.

  > **Note:** About how the MIN parameter is applied: For an optimal design, the program will allocate at least the minimum number of points per cell, even if it exceeds the number required to meet the desired accuracy. Of course, if the minimum requirement is greater than the cell population, the program will not specify a number greater than the population. For a fixed size sample, the program will first allocate the total number of available sample points across the cells according to the Neyman allocation. Then, it will increase the allocation in each cell to meet the minimum requirements, and remove the same number of points from the remaining cells according to selection rules that cause the least impact to overall precision.

- **HDn** — The commands HD1 and HD2 enable you to label the sample design outputs. Both heading commands are optional.

  There must be a space between the command name (HD1 or HD2) and your title. The title can be any combination of printable characters. It can be no longer than 76 characters.

```
DES OPT 5 90 19
HD1 CERTIFICATION TEST CASE #1 — 2 DIMENSIONAL DESIGN
HD2 DIM 1 SUMMER KWH DIM2 WINTER — DESIGN VARIABLE JULY KWH
```

Figure 8-2 Sample Design Environment File

For an Optimal Size Sample Analysis, with a Minimum Number of Points Per Cell Specified.

```
DES FIX 300
HD1 CERTIFICATION TEST CASE #2 — 2 DIMENSIONAL DESIGN
HD2 DIM 1 SUMMER KWH DIM2 WINTER — DESIGN VARIABLE JULY KWH
```

Figure 8-3 Sample Design Environment File for A FIXED Size Sample Analysis

Step 2: Run the Sample Design Program

Use the B320 Submit panel.

Step 3: Check Output

The Multidimensional Sample Design Program produces five outputs:

- **Multidimensional Sample Design Environment Report (TGB320-01)** — lists back the parameters you entered in the Environment File. Be sure to check this report to verify your input.
- **Multidimensional Sample Design Execution Log (TGB320-02)** — lists back the information from the Population Statistics File, along with messages about any errors the program encountered during processing.
- **Multidimensional Sample Design Report (TGB320-03)** — presents the results of the sample design. You will need this information for Step 5, and for Multidimensional Sample Selection.
- **Multidimensional Sample Design Summary Report (TGB320-04)** — lists information about the overall sample.
- **Relative Accuracy File (TGB52B.RAF)** — will be used as Sample Validation Environment File for Sample Validation Program (B510). This file contains a single command that identifies the required precision you specified in the Sample Design Environment File.

Step 4: Repeat the Preceding Steps for Each Additional Variable

You need to repeat the preceding steps for each additional usage variable in your design. As explained earlier, you must run the Multidimensional Sample Design Program to determine cell sizes for each usage variable in the multidimensional design. The only difference in each run is the contents of the Population Statistics File, which you must have created specifically for every dimension using the Multidimensional Population Analysis Program or a Oracle Utilities Analysis Program and prior load research data.

Step 5: Manually Combine Cells

In the preceding steps, you calculated the number of sample points per cell required to satisfy accuracy requirements for individual usage variables in the design. In this step, you will manually combine these sizes to create a single sample design that will simultaneously attain the confidence or accuracy criteria for all design variables simultaneously. To do this, you will compare, cell by cell, the number of sample points required for each usage variable, and select the largest number.

To facilitate this process, we recommend that you draw a grid similar to that shown in Table 8-1. List the cell numbers along the side, and the usage variables across the top. Add a column at the right for "Combined", and a row along the bottom for "Total". Under each variable, list the number of sample points required per cell. You will find this information listed under "Reallocated Sample Size" in the Sample Design Report (TGB320-02) — you should have one report for each usage variable. If any cell has a zero value, combine that cell with an adjacent cell in the same dimension.

Then, for each cell, select the largest number you listed in the row, and write that number under "Combined." These values are the number of sample points required to satisfy the confidence and accuracy requirements for all dimensions in the sample, and can be used as input to the Sample Selection Program (next chapter).

Table 8-1: To Combine Your Designs

	SAMPLE SIZE		
	JANUARY	**JULY**	**COMBINED**
Cell 1	91	79	91
Cell 2	30	10	30
Cell 3	15	40	40
Cell 4	66	58	66
TOTAL	202	187	227

Chapter 9

Selecting the Sample for a Single Dimensional Design

Once you have satisfactorily determined your single dimensional sample design (number of strata, strata breakpoints, and number of sample points per strata), you are ready to draw a list of actual customers for participation in the study.

Sample Selection Procedures

To select a list of customers for your sample, you will apply a **Sample Selection Procedure**. There are three methods of selecting sample populations available to you from the GUI. These are:

- Random Selection
- Systematic Sampling
- Centered Systematic Sampling

This chapter provides a detailed description of each of these Sample Selection methods.

Basically, each of the three Sample Selection procedures assigns each eligible customer in the Population Data File (.PDF) to its appropriate stratum and assigns a random number to each customer. Within each strata, the procedure picks the first "**n**" customers, where **n** is a number determined from your specifications in the Control File. The procedure then sorts the customers by strata assignment and assigned random number. Finally, the procedure produces a report of the selected customers, with descriptive information such as name and address, formatted to your specifications. Using the User Language Write Command in the Reporting Control File, you can also output the customers and descriptive information to a file for use by other software programs.

Each of the Sample Selection procedures produces two sets of statistics that will be needed for the Sample Validation Phase (**Chapter 11**: **Validating the Sample**): strata mean and standard deviation for the population, and strata mean and standard deviation for the selected sample. The program writes these statistics to the Population Statistics File and Sample Statistics File, respectively.

About Sample Alternates and Validation

You may wish to draw some number of alternate customers for your sample, since it is likely that not all customers selected will be willing or able to participate in the study. If you desire to select alternate customers for your sample, you can employ one of the following methods:

1. Multiple lists - include the LISTCOUNT command in the Control file.
2. Oversampling - increase the size of your sample selection for each strata (see "Step 5 Modify your Reporting Control File")

See Figure 9-1 for more information about selecting alternates.

About Version Numbers (Seed Numbers) and the Random Number Generator

To use the random number generator, you supply a version number, or "seed" that starts the propagation of a sequence of random numbers. The same version, or seed generates the same sequence of random numbers each time it is used. By specifying the same version seed numbers for the program's random number generator, you ensure that the program will draw the same customers each time.

For all three sampling methods, the random number is stored into the Population Data File, which is later sorted by that random number.

> **Note:** *If you are using the User Language to create your control file, it is recommended that the random number seeds be* **odd** *numbers.* Odd-number seeds generate the best results, for the following reason. Chains of random numbers will eventually repeat; the portion up to, but not including, the point where the random numbers begin to repeat is called a "period." Odd-number seeds yield longer periods than even-numbered seeds, and are therefore preferred.

Sample Selection

For Random sampling, Systematic sampling, and Centered Systematic sampling, the Sample Selection Procedure:

1. Assigns each eligible customer in the Population Data File (.PDF) to its appropriate stratum and gives the customer a random number.
2. Within each strata, the procedure picks the first "**n**" customers, where **n** is a number determined from your specifications in the Control File.
3. The procedure then sorts the customers by strata assignment and random number.
4. Finally, the procedure produces a report of the selected customers, with descriptive information such as name and address, formatted to your specifications. Using the User Language Write Command in the Reporting Control File, you can also output the customers and descriptive information to a file for use by other software programs.

If you are selecting a sample for a simple random design, follow the steps described in this chapter, specifying one stratum throughout. (If you are performing simple Random Selection without stratification, all customers are assigned to the same stratum.)

The Sample Selection Programs

The Sample Selection Procedure consists of three programs: Stratification, Sorting, and Reporting. The steps you take to execute these programs are summarized in the table below, and described in detail on the following pages.

SUMMARY
SELECTING THE SAMPLE FOR A SINGLE DIMENSIONAL SAMPLE DESIGN USING THE SAMPLE SELECTION PROGRAM (B410)

1. Verify or create the required inputs:
   - Population Data File (.PDF)
   - Record Definition File (.DEF) — Use the file created by B110.
   - Stratification Control File (TGB22A.CTL) —— strata breakpoints, random number seeds, and instructions for creating the Population Statistics File, based on information from the Single Dimensional Sample Design Report produced by B310.
   - Although the reference numbers are the same (TGB22A.CTL), this is not the same Control File used by the Single or Multidimensional Population Analysis programs, or that described in Step 1D below.
   - Sort Control File — sort parameters; created automatically from the Sampling Parameters File specifications.
   - Reporting Control File (.RCF) — definition of the sample report (number of customers, content, and format) and instructions for creating the Sample Statistics File.
2. Submit the job (Sample Selection Procedure — B410).
3. Check output.

   From Stratification Program
   - Population Analysis Execution Log (TGB220-01)
   - Selection File (TGB222)[a] — unsorted Population File records (temporary data file)
   - Population Statistics File (.PSF)**

   From Sort Program
   - Sorted Selection File* (temporary data file)
   - Report of sorting parameters

   From Reporting Program
   - Population Analysis Execution Log (TGB220-01) Summary Selection Report (TGB220-01)
   - Sample Statistics File (.SSF)[b]
   - Table of Cell Entries Report — printed version of TGB223 in a slightly modified format.

Optional. If you wish to draw alternates:

1. Validate your sample selection using the Sample Validation Program (B520). If the sample passes, go to Step 5. If not, modify the Stratification Control File with a new set of random number seeds and resubmit the Sample Selection Procedure (B410).
2. Modify the Reporting Control File to specify an allowance for alternates.
3. Resubmit the Sample Selection Procedure (B410) and check output. The sample design, selection, and validation process is now complete.

a. ***Warning about space requirements!*** Each time you run the job, the programs create two temporary working copies of a subset of the Population Data File — the Selection File and a sorted version of the Selection File (the number of records in these files is determined by a cutoff value supplied in the Stratification Control File). These files tend to be very large. Also, the Sorting Program requires some working space — typically three times the size of the original Selection File. Depending upon the size of your file, you may need to ensure there is sufficient space in your job directory.

b. ***Warning about Statistics Files!*** Two programs in the Sample Selection Procedure produce files that are required inputs to Sample Validation (**Chapter 11**: **Validating the Sample**). The Stratification Program outputs the *Population Statistics File* (.PSF), and the Reporting Program outputs the *Sample Statistics File* (.SSF).

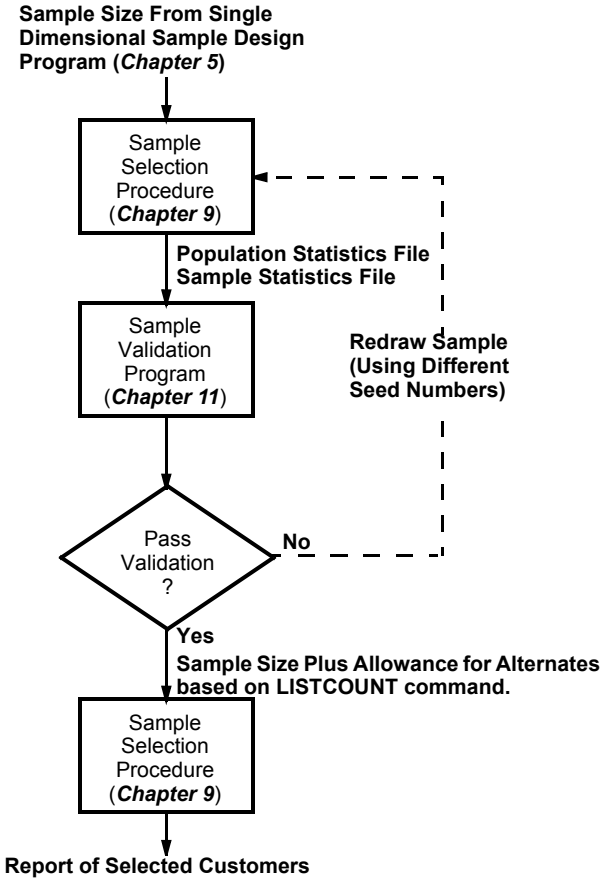


Figure 9-1 Selecting and Validating a Single Dimensional Sample with Alternates

Step 1A: Verify the Population Data File (.PDF)

Verify that the Population Data File (.PDF) has been properly set up and is available. It should be available for selection in the submit panel under the Population Data file dropdown. If you have not already created the Population Data File (.PDF), see Chapter Three.

Step 1B: Verify the Record Definition File (.DEF)

Verify that the Record Definition File (.DEF) has been properly set up. It should be available for selection in the submit panel under the Record Definition file dropdown. If you have not already created the Record Definition File (.DEF), see Chapter Three.

Step 1C: Create the Stratification Control File (TGB22A.CTL)

The Stratification Program selects eligible customers from the .PDF file; assigns each customer to a stratum based on the customer's value for the usage variable and the strata breakpoints; and assigns each customer a random number. The resulting output is the "Selection File" — a subset of the Population Data File, identical in format except that the fields reserved for the strata and random numbers now contain assigned values. Only the number of customers you request are output to the Selection File. This file is used as input to the other two programs in the Sample Selection Procedure.

Before the program can execute these steps, you will need to supply the following information in the Stratification Control File:

- **Definition of eligible customers** (that is, the target population) — this is necessary only if the target population is a subset of the Population Data File.
- **Strata breakpoints** — the lower boundaries for the strata you determined in the Sample Design Phase. Refer to your Sample Design Report (TGB310-10) for these values; guidelines for finding the values on the report are provided in **Chapter 5**: **Stratifying the Population and Determining Sample Size for a Single Dimension**.
- **Version numbers (Random number seeds) and Group Sizes (Cutoff values)** — for each stratum, you must supply a version number (seed) for the program's random number generator, and a group size (cutoff value) used to determine the number of customer records output to the Selection File (thereby saving space and processing time).

Each of these elements is explained in more detail below.

Create the Stratification Control File

The commands for creating stratification control file are as follows:

SELect	field = value
USAge	usageVariable
BREakpoints	strata1Boundary, strata2Boundary, … strataNBoundary; INF
VERsion	strata1version, strata2version, … strataNversion
GROup_size	strata1groupsize, strata2groupsize, … strataNgroupsize
LIStcount	x

SELect	field = value

SELECT (optional) — SELECT is a test statement for defining eligible customers. For example, to process records whose RATE code is 'SGS' only: SELECT RATE = 'SGS'

USAge	usageVariable

USAGE — Defines the usage, or design variable. Specify the name of the usage variable you would like to use. For example:

USAGE JAN

BREakpoints	strata1Boundary, strata2Boundary, … strataNBoundary; INF

BREAKPOINTS — Defines the lower stratum breakpoints. The last value in the list must be INF (infinity). Comma separated.

VERsion	strata1version, strata2version, … strataNversion
GROup_size	strata1groupsize, strata2groupsize, … strataNgroupsize

VERSION — Specify the version # (1 – 100) for each stratum. Different version numbers result in a different a set of customers drawn.

GROUP_SIZE — For each stratum, specify the size of each group from which to draw customers from. Group size can be a numeric value or percentage depending on the Selection Method used:

- Simple Random Selection – For each stratum, specify a value between 0 – 1, or a percentage value between 0% - 100%. 100% means all customers in the stratum are eligible for random selection.
- Systematic Selection – For each stratum, specify the size of each group from which to select customers from. A group size of 10 means to randomly select 1 customer out of every group of 10.
- Systematic Centered – For each stratum, specify the size of each group from which to select customers from. A group size of 10 means to select the 5th customer out of every group of 10.

LIStcount	x

LISTCOUNT (Optional) — Indicate the total number of lists to be generated. x must be an integer from 1 to 8. The default is 1, so if LISTCOUNT is not included, one list will be generated.

Sample B410 Stratification Control File:

USAGE	JAN
BREAKPOINTS	4000, 10000, INF
VERSION	1, 1, 1
GROUP_SIZE	1, 1, 1
LISTCOUNT	1

Create the Stratification Control File with User Language

You can optionally compose the Stratification Control File with User Language statements. If you already have constructed your analysis control file, you can skip this section. The "User Language" allows for more customization, but keep in mind it is also more complex.

You build the Stratification Control File with User Language statements similar to those you employed for the Population Analysis Control File (**Chapter 4**: **Analyzing the Population Frequency Distribution**). However, instead of counting customers within usage intervals, the statements you create here assign customers to strata based on the usage breakpoints you determined in **Chapter 5**: **Stratifying the Population and Determining Sample Size for a Single Dimension**.

A template for creating the Stratification Control File is shown in Figure 9-2. To create your own Stratification Control File, you may find it easiest to simply modify the example provided in Figure 9-3.

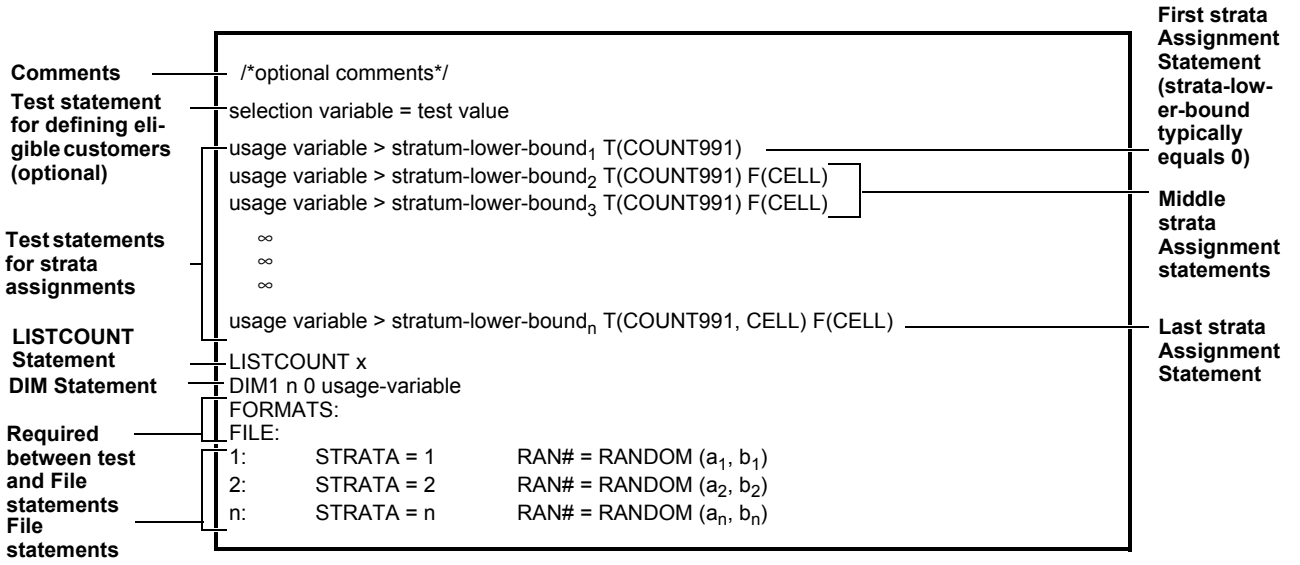


Figure 9-2 Stratification Control File "Templates"

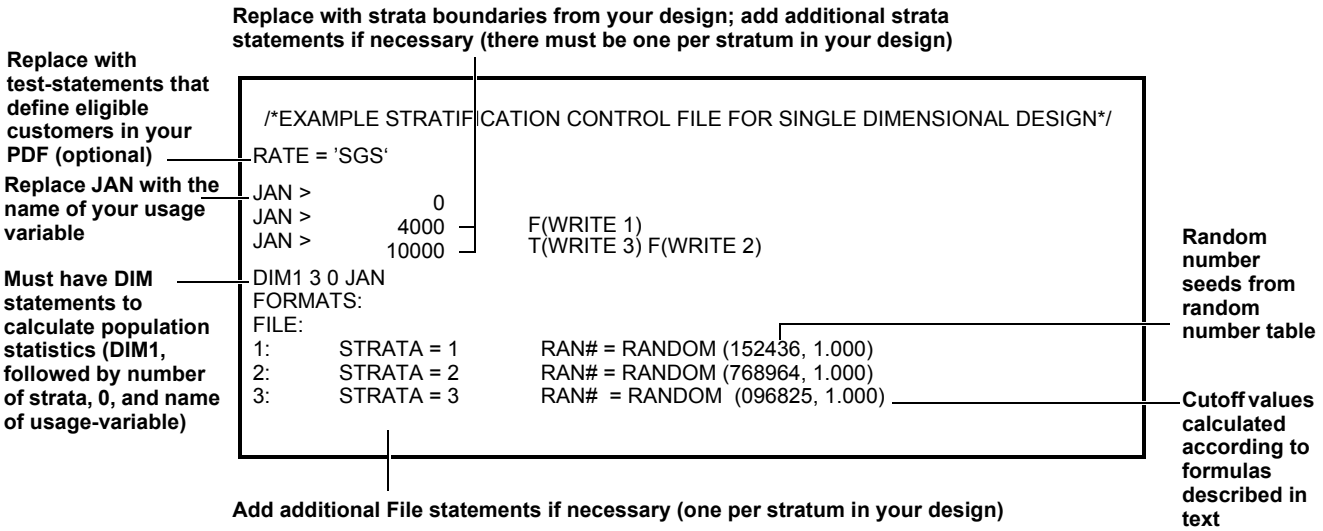


Figure 9-3 Example Stratification Control File

You may be able to create your own file by simply modifying this example.

As you can see from the template and example, a typical Stratification Control File is composed of a few basic elements:

1. **Comments** — Optional notes that do not affect processing. You must enclose any comments between these symbols: /* and */. Do not split comments across lines, unless you begin and end each line with the slash/asterisk symbols. Do not enter the slash/asterisk symbols in columns 1 and 2 of the file; otherwise, you may put comments anywhere in the file.

2. **Test statements** — Test statements have two purposes in a typical Stratification Control File: 1) to select a subpopulation of the PDF as eligible for inclusion in the sample; and 2) to assign customers to the appropriate strata.

   Test statements for defining eligible customers: You use exactly the same test statements you used in the Population Analysis Control File to define eligible customers (the *target population*). These statements are required only if the target population is a *subset* of the PDF.

   Test statements for strata assignments: You must create a series of test statements that compare each customer's value for the usage variable to the strata boundaries you determined with the Sample Design Program. The typical format for strata Assignment statements is:

   usage-variable > stratum-lower-bound$_{1...n}$ T(COUNT991) F(CELL)

   where:

- **usage-variable** is the name of your design variable, spelled exactly as it appears in your Record Definition File;
- **stratum-lower-bound** is the lower boundary of the stratum, as determined in the Sample Design phase;
- **n** equals the total number of strata in your design.

   Using the logic you construct, the program compares each customer's value for the usage variable to a series of strata boundaries. The COUNT991 counts the number of times a customer passes a test-statement, until it finally fails and falls into a bucket, i.e., the stratum in which it belongs. The number of times the counter was triggered equals the stratum number for that customer. When a customer fails a test-statement, it also triggers the Cell option. The Cell option works with the DIM Statement to compute population statistics for each cell, and refers the program to a matching File Statement that assigns a random number.

   The first strata Assignment Statement is typically:

- **usage variable > 0**, which eliminates zero usage customers from the selection process.

   The last statement is typically:

- **usage variable > stratum-lower-bound$_n$ T (COUNT991, CELL) F (CELL)**

   where **n** is the number representing the last stratum in your design (i.e., equal to the total number of strata).

3. **LISTCOUNT Statement**— A statement that indicates the total number of lists to be generated. These lists can be treated as a primary list and x-1 alternate lists. The LISTCOUNT statement should be placed immediately before the DIM statement if there is one, or immediately before the FORMATS statement if the DIM statement is not present. Use this format to construct your LISTCOUNT Statement for sample selection:

   LISTCOUNT x

   where:

- **LISTCOUNT** indicates that you are creating more than one list.
- **x** indicates the total number of lists to be generated and must be an integer from 1 to 8. The default is 1, so if LISTCOUNT is not included, one list will be generated.

   Notes:

a. The lists generated by Single Dimension Selection B410 will be named TGB413.LST1 thru TGB413.LSTx where x is the number of lists generated. The lists generated by Multi Dimension Selection B420 will be named TGB423.LST1 thru TGB423.LSTx.

b. If a DIM statement and STATS statements are specified in the Report Control File then a Sampling Statistics File will be generated for each list. These will be named with suffixes .SSF1 thru .SSFx where x is the number of lists generated.

c. The specifications coded in the Sample Selection control file should be based on the user's requirements for the primary list only. The Sample Selection program will use those specifications, combined with the value specified for LISTCOUNT, to select enough customers to fill all the lists requested, as long as enough eligible customers exist.

d. The lists will be filled in numerical order, so if the system runs out of eligible customers the higher numbered lists will be shorter than the lower numbered lists.

5. **DIM Statement** — You must include one DIM Statement in your Stratification Control File. This statement provides information that the program uses to calculate the population statistics required for Sample Validation. The format of the statement was designed to accommodate both single- and multidimensional samples, so there are more features available than you will need here. Use this format to construct your DIM Statement for a single dimensional sample selection:

   DIM1n b usage-variable

   where:

- **DIM1** indicates that you are creating a single dimensional design.

  Do **not** place spaces between any of these characters. Correct: DIM1. Incorrect: DIM 1.

- **n** is the number of strata in your design; must be an integer between 2 and 9, inclusive.
- **b** is the number of special cells in your design; specify 0.
- **usage-variable** is the name of the design variable in the Population Data File for which the statistics (mean and stratum standard deviation) are to be computed in the program run. Must be the same name you used in the test-statements for strata assignments.

  You may place one or more blanks between the parameters in the statement. Correct example: **DIM1 4 0 JAN**

  For example, let's say that you have 3 strata and "**JUL**" is the name of the variable in the Population Data File that you want to use as the design variable in this program run. You would construct the DIM Statement as follows:

  DIM1 3 0 JUL.

6. You must place the keywords "**FORMATS:**" and "**FILE:**" between the DIM Statement and the File statements, as shown in the examples.

7. **File statements** — A File Statement is a special kind of format statement that controls the content and organization of an output file (in this case, the Selection File). In a typical Stratification Control File, File statements add the strata numbers and random numbers to selected customer records in a copy of the Population Data File. For each Write clause in the test statements, you must have a matching File Statement.

   For Random Sampling, the file statement should be set up in the following format:

   $y_{1...n:}$ STRATA = $y_{1...n}$ RAN# = RANDOM(a,b)

   where:

- **y** is a sequential number beginning with 1. It is also the number in the associated Write clause, and effectively the strata number.

- **a** is a random number seed that the program uses to initialize the random number process. **Specify the seed as a 6-digit, odd integer between 100000 and 999999. Be sure to use a unique seed number for each Strata Statement**. You can obtain these seeds from a table of random numbers. ***Note:*** Each seed number generates the same sequence every time. This is useful should you need to re-create your process.

- **b** is a cutoff value that limits the number of records that the program must sort and report, saving processing time. (***Note:*** The use of a cutoff value *does not* mean Systematic Sampling. Instead, this process assigns all customers a random number and eliminates those with a low probability of selection.)

  Specify the cutoff value as a real number between 0.0 and 1.0. If you specify a cutoff value of 1.0, every qualifying customer in the PDF will be selected for output in the Selection File.

  The cutoff value is used to indicate the percentage of your population that is to be used in the sample (i.e., 0.33 = 1/3 of the population). The random numbers are always a real number between 0 and 1. If a customer's random number is less than the cutoff value, that customer is included in the sample. Therefore, a cutoff value of 0.33 yields a sample population that consists of approximately 1/3 of the population; the other 2/3 are discarded because their random numbers are above the cutoff value.

  How to calculate cutoff values: For large target populations (that is, greater than 30) you can calculate the cutoff value for a particular stratum using this formula: ***c=n/N***

  where:

- **n** is the required sample size for the stratum, including an allowance for alternates. You will typically use a value that is five times greater than the number of sample points required for the stratum, as calculated by the Sample Design Program. For example, if your design called for 20 sample points for the stratum, your value for **n** would be 100.

- **N** is the stratum population size.

  For small samples, the cutoff value must be adjusted to ensure an adequate selection. In this case the appropriate formula is:

  $c = [n + 2 * SQRT (n - n^2 / N)] / N$

  where ***n*** and ***N*** are the same as described above. Typical cutoff values for small populations and sample sizes are presented in Table 9-1.

  .

Table 9-1: Cutoff Values for Small Populations

	POPULATION SIZE (N)						
SAMPLE SIZE (n)	**50**	**100**	**250**	**500**	**1000**	**5000**	**100000**
10	0.3131	0.1600	0.0648	0.0325	0.0163	0.0033	0.0002
25	0.6414	0.3366	0.1379	0.0695	0.0340	0.0070	0.0003
60	1.0000	0.6000	0.2506	0.1260	0.0638	0.0128	0.0006
100		1.0000	0.4620	0.2358	0.1190	0.0240	0.0012
250			1.0000	0.5447	0.2774	0.0562	0.0028
500				1.0000	0.5316	0.1085	0.0054

Systematic Sampling selects the **R**th point from an ordered population, and every **k**th point thereafter. **k** is determined by dividing **N** by **n**, where **N** is the entire eligible population and **n** is the desired sample size. **R** is a random number between 1 and **k**, inclusive.

In effect, the population is divided into **n** groups, each containing **k** customers (with the last group possibly containing fewer than **k** customers). One customer from each group is

chosen. For example, if the population consisted of 300 customers (**N**) and was divided into 30 groups, each group would consist of 10 customers, and selecting one customer from each group would yield a sampling population of 30 customers (**n**). When selecting one customer from each group, the Systematic Sampling method always selects the customer with the same relative position within each group. In other words, Systematic Sampling takes a random number between 1 and **k**, where **k** is the size of each group, and selects the customer in each group whose position within the group is equal to that random number. So if, for example, the group size is 10 and the random number selected is 8, the 8th customer from each group is selected for inclusion in the sample population.

Unlike simple Random Sampling, Systematic Sampling forces an even distribution across the entire population from which sample points are drawn, providing a better representation of the population.

For Systematic Sampling, the File Statement is similar to that for Random Sampling, with this difference:

- **b** is used to specify the size (**k**) of each of the groups into which the population is to be divided for selection.

The following File statements might be used for Systematic Sampling:

FILE:

1: STRATA = 1 RAN# = RANDOM (152433, 5)

2: STRATA = 2 RAN# = RANDOM (1768961, 10)

3: STRATA = 3 RAN# = RANDOM (596825, 7)

In this case, one customer would be selected from each group of 5 in the stratum 1 population, one from each group of 10 in stratum 2, and one from each group of 7 in stratum 3.

Centered Systematic Sampling is a refinement of the Systematic Sampling technique. In this case, the random number **R** is equal to **k / 2**. Unlike the Systematic Sampling method, Centered Systematic Sampling selects the middle customer of each group, rather than generating a random number to select a customer from each group.

The same File statements shown above for Systematic Sampling could be used for Centered Systematic Sampling, the difference being which customer would be selected from each group. For example, in stratum 1 the third customer from each group would be selected.

Note: For all three sampling methods, the random number is stored in its slot in the population record, even though the third method, Centered Systematic Sampling, does not use the random number. The population is later sorted by that random number, so random number seeds are still needed regardless of the sampling method used.

Step 1D: The Sort Control File is Created

The Sort Control File is created automatically, based on the values supplied in the Sampling Parameters File.

Step 1E: Create the Reporting Control File (.RCF)

During the Reporting Phase, the program creates a report of a user-defined number of selected customers, formatted to your specifications. The report typically includes (but is not limited to):

- Customer identifier or account number.
- Selection variables or other qualitative characteristics such as service company, rate class, and SIC code. If you are eventually going to perform systematic replacement, you need to supply user-defined variables to identify alternates with similar characteristics (see **Step 5: Modify your Reporting Control File** on page 9-17 for more information about alternates).
- Additional data for installation or mailing purposes, such as name, address, and telephone number.
- Stratum number (use one stratum for all customers if your sample will not be stratified.

The program also calculates statistics for the selected sample, and outputs the statistics to the Sample Statistics File (.SSF). These statistics are required for Sample Validation.

Along with the Population Data File (.PDF) and the Record Definition File (.DEF), the Reporting Phase requires a **Reporting Control File (.RCF)**.

You use the Reporting Control File to define the content and format of the selection report, including the number of candidates to be reported within each stratum, and to output the required statistics. For the first program run, ***be sure to specify just the required number of sample points per strata***, so that the statistics created for validation will reflect the actual sample size. (You will have an opportunity to specify some number of alternates later in the process.)

Create the Reporting Control File

The Reporting Control File is created with the following commands:

USAge	usageVariable
SAMple_size	stratum1SampleSize, stratum2SampleSize … stratumNSampleSize
PRInt	printVariable1, printVariable2 … printVariableN
DELimiter	BLAnk
	COMma

USAge	usageVariable

USAGE — Defines the usage, or design variable. Specify the name of the usage variable you would like to use. For example:

```
USAGE JAN
```

SAMple_size	stratum1SampleSize, stratum2SampleSize … stratumNSampleSize

SAMPLE_SIZE — List the sample size for each stratum. Comma separated.

PRINT — Provide the list of fields to print for each selected customer in the final selection

PRInt	printVariable1, printVariable2 … printVariableN

report. Comma separated.

DELIMITER	BLAnk

DELIMITER – (optional) controls the formatting of the output list file(s). **BLAnk** indicates to produce a space delimited output file. This is the default. **COMma** produces a comma separated output list file.

Sample RCF File

USAGE	JAN
SAMPLE_SIZE	20, 22, 124
PRINT	STRATA, CUSTID, NAME, ADDRESS

Create the Reporting Control File with User Language

You can optionally compose the Reporting Control File with User Language statements. If you already have constructed your Reporting Control File, you can skip this section. The "User Language" allows for more customization, but keep in mind it is also more complex.

You create the Reporting Control File with User Language statements. An example Reporting Control File is shown in Figure 9-5. To get started, you may be able to simply modify the example Control File shown in Figure 9-4. It is likely however, that you will wish to create a more detailed report tailored to your specific needs. The User Language is very flexible and will enable you to output any data in your Population Data File, formatted according to your own design. **To produce more complex or customized reports, you may find it necessary to refer to the User Language instructions provided in Appendix A**: **The User Language**.

Important Note About DIM Statement: To generate the Sample Statistics File required for Sample Validation, you must include the DIM Statement and STATS option, as shown in Figure 9-4 and Figure 9-5 (use the same DIM Statement you created for the Stratification Control File).

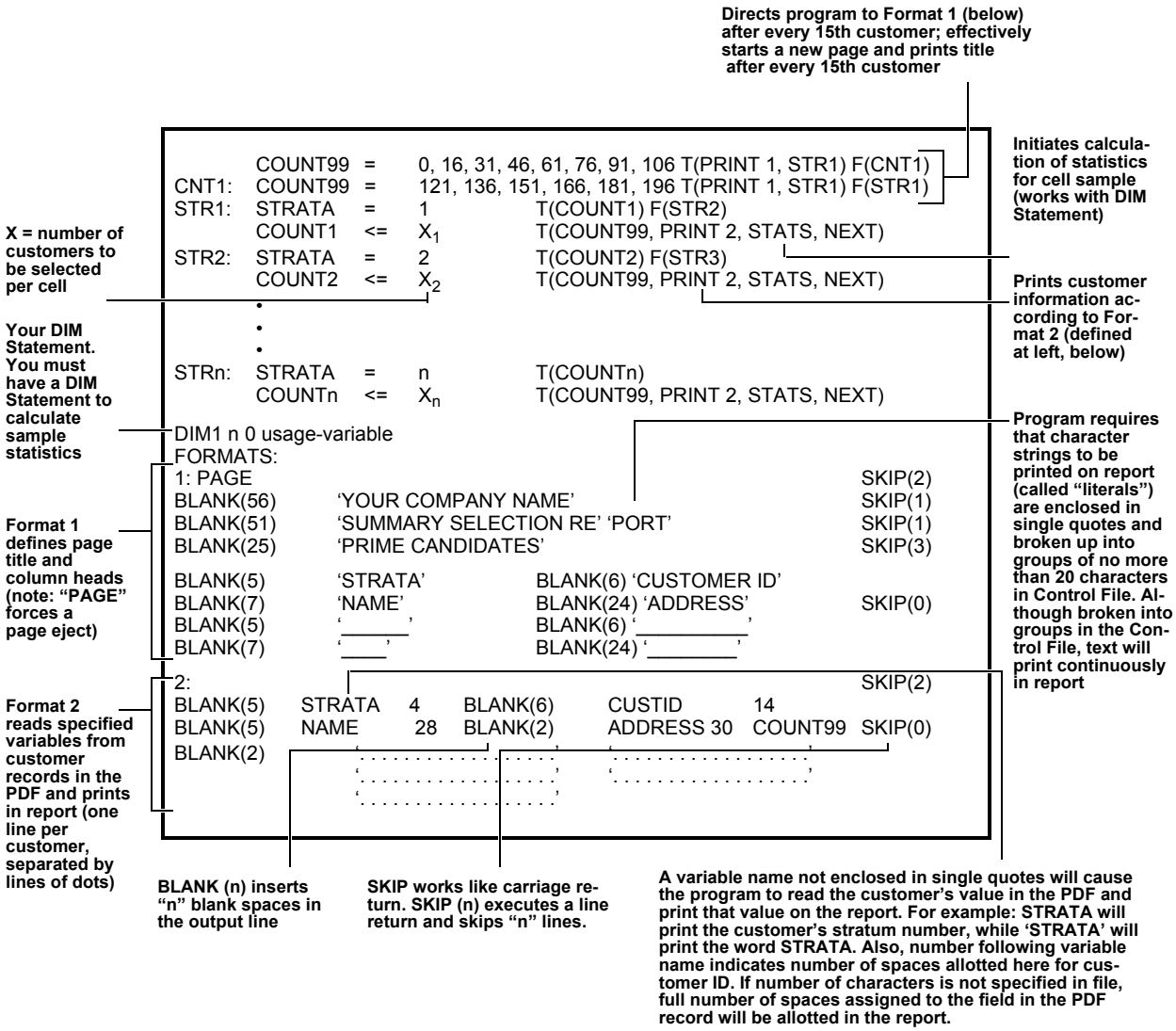


Figure 9-4 Reporting Control File "Template"

The template in Figure 9-4 produces a Sample Statistics File. It assumes that variable names in PDF for customer ID, name, and address are CUSTID, NAME, and ADDRESS, respectively.

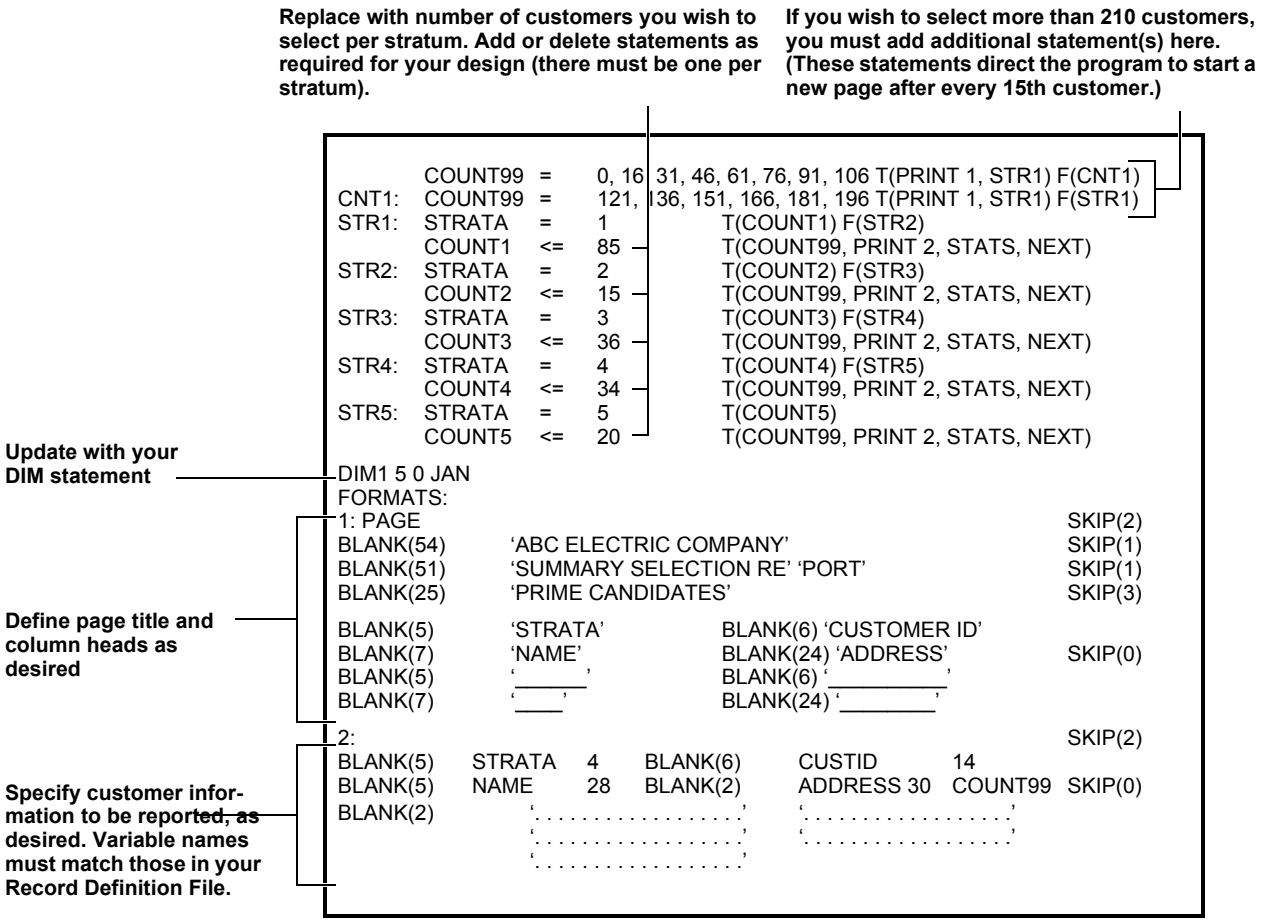


Figure 9-5 Example Reporting Control File

This example was used to produce a Sample Statistics File. It illustrates how you can use the template to create your own file and report.

Step 2: Submit the Sample Selection Procedure

Use the B410 **Submit Panel**.

The first page of the B410 **Submit Panel** contains the following fields:

- Control File - the Stratification Control File you created in Step 1C (e.g., TGB22A.CTL)
- Report Control File - the .RCF Reporting Control File you created in Step 1E
- Record Definition File - the .DEF file from the COMMON\DATA directory
- Population Data File - the .PDF file from the COMMON\DATA directory
- Sample Processing Mode - select either TRIAL or PRODuction. See **File Placement** on page 1-5 for more information.
- Population Statistics File - the .PSF file
- Sample Statistics File - the .SSF file.

Fill in or select values for these fields, and then click on the page "ear" (in the above illustration, the white triangle near the Close button) to display the second page of the B410 **Submit Panel**.

The second page of the B410 **Submit Panel** contains the following fields:

- Sampling Parameter File - the .SPF file
- Selection Method: select RANDOM, SYSTEMATIC, or CENTERED.

Fill in or select values for these fields, and then click on the page "ear" (in the above illustration, the white triangle above the Sequencer tab) to return to the first page of the B410 **Submit Panel**.

Click the **Submit** button to submit the Sample Selection job.

Step 3: Check Output

Each program in the Sample Selection Procedure produces its own set of outputs:

Stratification Program

- **Population Analysis Execution Log (TGB220-01)** — Contains a copy of the Stratification Control File and any processing messages.
- **Selection File (TGB222)** — Copy of the PDF (SORTED.PDF), with the reserved fields STRATA and RAN# filled in. Only the number of customers you requested in the Stratification Control File (via the cutoff value) are output to the Selection File. This file is used as input to the program sorting. Be sure to allocate space for this file.
- **Population Statistics File (.PSF)** — Strata means and standard deviations for the population. You will use this file as input to the Sample Validation Program.

Reporting Program

- **Population Analysis Execution Log (TGB220-01)** — Lists the Reporting Control File, any processing messages, and the desired customer selection report.
- **Sample Statistics File (.SSF)** — Strata means and standard deviations for the sample. You will use this file as input to the Sample Validation Program.
- **Table of Cell Entries Report** — Printed version of the Sample Statistics File in a slightly modified format.

Step 4: Validate your Sample Selection

You can validate your sample selection using the sample validation procedure. Before drawing alternate sample points, it is important that you first validate your initial sample using the statistics produced by the Sample Selection Procedure. Follow the instructions provided in **Chapter 11**: **Validating the Sample**.

- *If the sample passes validation*, go to Step 5 below.

If the sample fails validation, replace the version number (random number seeds if you are using User Language) in your Stratification Control File (TGB22A) with a completely new set, and rerun the Sample Selection Procedure. Repeat this process until you arrive at a satisfactory sample selection, then proceed to Step 2 below.

Step 5: Modify your Reporting Control File

You can modify your reporting control file (TGB22A.RCF) to specify an allowance for alternates. For your final sample selection, you will typically request the number of sample points specified by the sample design program *plus* an allowance for alternates. Typically, not all customers selected will be willing or able to participate in the sample, so you need to add some allowance that ensures getting the number of sample points you require. The pool of alternates will be selected in sort order, based on the random number assignment.

In your Reporting Control File, update the number of customers to be selected per stratum. The numbers you specify will depend upon your criteria for selecting alternates:

- *If you intend to replace primary sample points with the next random selection,* we recommend that you select two to three times the number of customers required to meet the desired accuracy.
- *If you must select alternates with like characteristics (that is, systematic replacement)*, you will need a large pool from which to draw. We recommend that you select five times the number of required sample points. When defining the Reporting Control File, keep in mind also that you must include the pertinent selection variables in your report, so that you can identify like customers. For example, a criterion for drawing alternates might be that they share the same meter reading route, in which case you would have to identify each customer's route on the report.

  Note: If you request more customers than have been assigned to a stratum, the program will simply select all members of that stratum. The program will not issue an error message.

Step 6: Resubmit the Sample Selection Procedure (B410) and Check Output

You have now completed the sample design, selection, and validation process. The result should be a sample that accurately mirrors the targeted population and meets all design criteria.

Chapter 10

Selecting the Sample for a Multidimensional Design

Once you have satisfactorily determined your multidimensional sample design (number of sample points per cell), you are ready to draw a list of actual customers for participation in the study.

Sample Selection Procedures

To select a list of customers for your sample, you will apply a **Multidimensional Sample Selection Procedure**. There are three methods of selecting sample populations available to you from the GUI. These are:

- Random Multidimensional Selection
- Systematic Multidimensional Sampling
- Centered Systematic Multidimensional Sampling.

This chapter provides a detailed description of each of these Multidimensional Sample Selection methods.

In this process, you will use a **Multidimensional Sample Selection Procedure** to choose a random sample of customers from the Population Data File according to your sample design criteria. Each of the three Sample Selection procedures divides the eligible customers into cells (according to the cell numbers assigned by the Multidimensional Population Analysis Program) and assigns each candidate a random number. Next, it sorts the customers by cell assignment and random number. Within each cell, the procedure picks the first "n" customers, where **n** is determined from your specifications in the Control File. Finally, the procedure produces a report of the selected customers with descriptive information such as name and address, formatted to your specifications. Using the User Language Write Command in the Reporting Control File, you can also output the customers and descriptive information to a file for use by other software programs.

In addition, each of the Multidimensional Sample Selection procedures produce a set of sample statistics that will be required for the Sample Validation Phase (**Chapter 11**: **Validating the Sample**): strata mean and standard deviation for the selected sample. For that reason, ***if you intend to validate your sample using the Sample Validation Procedure, you must perform the steps described in this chapter once for each usage variable in your design.***

About Sample Alternates and Validation

You may wish to draw some number of alternate customers for your sample, since it is likely that not all customers selected will be willing or able to participate in the study. If so, include the LISTCOUNT command in the Environment file.

By specifying the same version numbers (seed) for the program's random number generator, you ensure that the program will draw the same customers each time.

See Figure 10-1 for more information about selecting alternates.

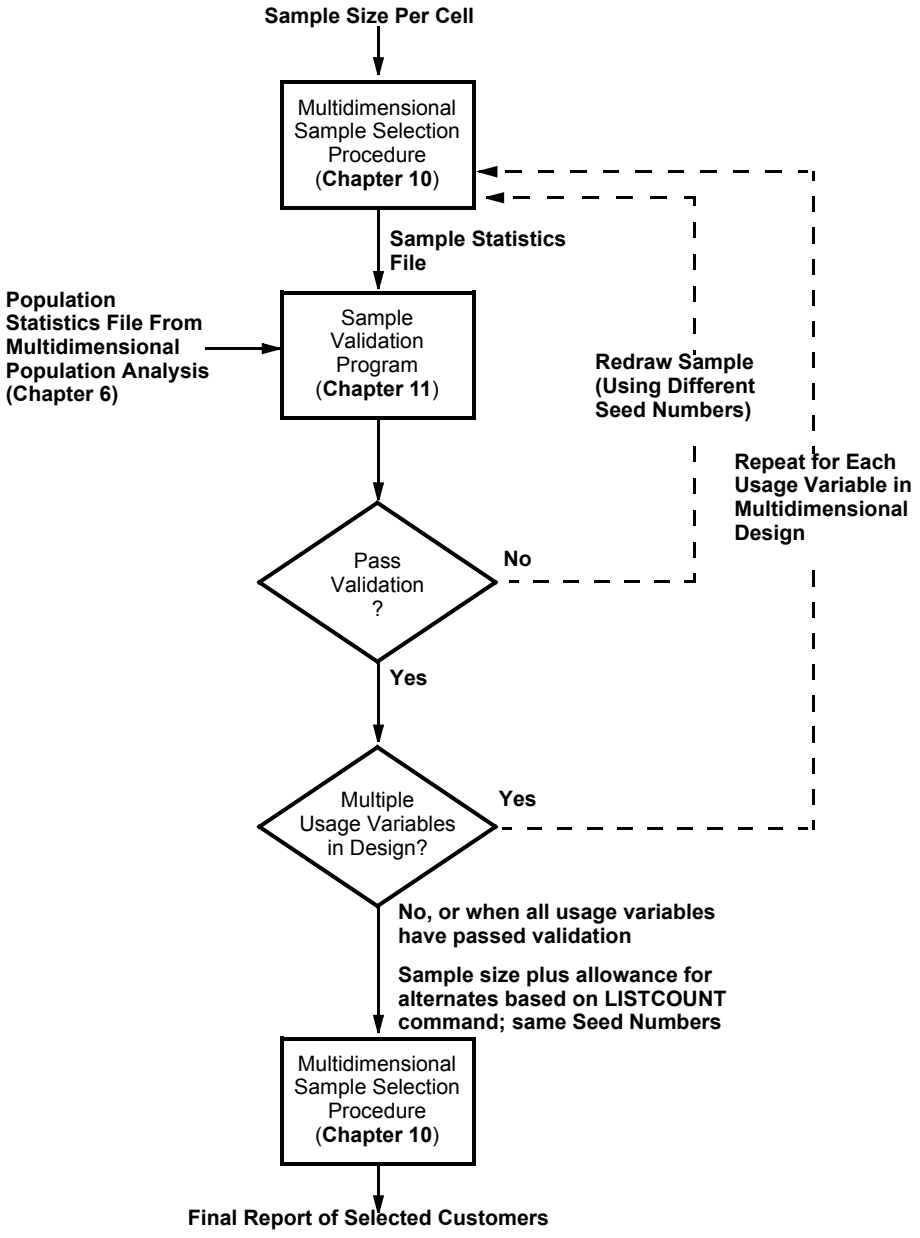


Figure 10-1 Selecting and Validating a Multidimensional Sample with Alternates

About Version Numbers (Seed Numbers) and the Random Number Generator

To use the random number generator, the user supplies a version number (or "seed" in the User Language) that starts the propagation of a sequence of random numbers. The same number generates the same sequence of random numbers each time it is used.

For all three sampling methods, the random number is stored into the Population Data File, which is later sorted by that random number.

Note: *If you are using the User Language, it is recommended that the random number seeds be* ***odd*** *numbers.* Odd-number seeds generate the best results, for the following reason. Chains of random numbers will eventually repeat; the portion up to, but not including, the point where the random numbers begin to repeat is called a "period." Odd-number seeds yield longer periods than even-numbered seeds, and are therefore preferred.

The Sample Selection Procedure consists of three programs: *Stratification, Sorting, and Reporting.* The steps you take to execute these programs are summarized below, and described in detail on the following pages.

SUMMARY
SELECTING THE SAMPLE FOR A MULTIDIMENSIONAL DESIGN USING THE SAMPLE SELECTION PROCEDURE (B420)

1. Verify or create the required inputs:
   - New Scored Population Data File (.PDS) — Use version created by the Multidimensional Population Analysis Program, e.g., .PDS from B220.
   - Record Definition File (.DEF) — Use the .DEF created by B110.
   - Stratification Control File (TGB22A) — primarily, random number seeds for each cell.

     Note: Although the reference numbers are the same (TGB22A), this is not the Control File used by Single or Multidimensional Population Analysis programs.
   - Reporting Control File (.RCF) — definition of the sample report (number of customers, content, and format) and instructions for creating the Sample Statistics File.
2. Submit the job (Sample Selection Procedure — B420).
3. Check output.

   From Stratification Program

   — Population Analysis Execution Log (TGB220-01)

   From Sort Program

   — Sorted Selection File (temporary data file)

   From Reporting Program

   — Population Analysis Execution Log (TGB220-01) Summary Selection Report (TGB220-01)

   — Sample Statistics File (.SSF)[a]

   — Table of Cell Entries Report — printed version of .SSF in a slightly modified format.

4. *Optional — only if your design has more than one usage variable AND you intend to validate your sample* — Rerun Sample Selection to calculate sample statistics for each additional usage variable in your design. **The only modifications required before rerunning Sample Selection are to change the USAGE Statement (DIM Statement if you are using the User Language) in the Reporting Control File and to assign a new dataset name to the Sample Statistics File.**

SUMMARY
SELECTING THE SAMPLE FOR A MULTIDIMENSIONAL DESIGN USING THE SAMPLE SELECTION PROCEDURE (B420)

Optional: If you wish to draw alternates by oversampling:

1. Validate your sample selection using the Sample Validation Program (B520). If the sample passes validation, go to Step 6. If the sample fails validation, modify the Stratification Control File with a new set of random number seeds, and resubmit the Sample Selection Procedure (B420). *Repeat for each usage variable in your design.*
2. When the sample selection has passed validation for all usage variables, modify the Reporting Control File to specify an allowance for alternates.
3. Resubmit the Sample Selection Procedure (B420) and check output. The sample design, selection, and validation process is now complete.

a. ***Warning about Sample Statistics Files!*** The Sample Statistics File output by the reporting Program is a required input to Sample Validation. *You must create one Sample Statistics File for each usage variable in your design, if you intend to perform validation.*

Step 1A: Verify that the Scored Population Data File is available (.PDS)

Verify that the Scored Population Data File (PDS) created by the Multidimensional Population Analysis Program is available. The Scored Population Data File should be available for selection in the Scored Population Data File drop down. In the scored population data file, each eligible member of the population has been assigned a cell number in the STRATA field.

Step 1B: Verify that the Record Definition File is available (.DEF)

Verify that the Record Definition File (.DEF) has been properly set up. It should be available for selection in the submit panel under the Record Definition file dropdown. If you have not already created the Record Definition File (.DEF), see Chapter Three.

Step 1C: Create the Stratification Control File (TGB42A.CTL)

The Stratification Program assigns a random number to each customer in the scored PDF (PDS), based on its cell assignment. The resulting output is the "Selection File" — another version of the Population Data File, identical in format except that the field reserved for the random number now contains an assigned value. Only the number of customers you request are output to the Selection File. This file is used as input to the other two programs in the Sample Selection Procedure.

Before the program can execute these steps, you will need to supply a Stratification Control File. In this file, you must supply version numbers (seed numbers) for the program's random number generator, and Group sizes (cutoff values). The Group sizes (cutoff values) specify the number of customer records output to the Selection File, thereby saving space and processing time. (Note: The use of a Group Size (cutoff value) does not mean systematic sampling. Instead, this process assigns all customers a random number and eliminates those with a low probability of selection.)

Create the Stratification Control File

The commands for creating stratification control file are as follows:

VERsion	strata1version, strata2version, … strataNversion
GROup_size	strata1groupsize, strata2groupsize, … strataNgroupsize
LIStcount	x

VERsion	strata1version, strata2version, … strataNversion
GROup_size	strata1groupsize, strata2groupsize, … strataNgroupsize

VERSION – Specify the version # (1 – 100) for each stratum. Different version numbers result in a different a set of customers drawn.

GROUP_SIZE – For each stratum, specify the size of each group from which to draw customers from. Group size can be a numeric value or percentage depending on the Selection Method used:

- Simple Random Selection – For each stratum, specify a value between 0 – 1, or a percentage value between 0% - 100%. 100% means all customers in the stratum are eligible for random selection.
- Systematic Selection – For each stratum, specify the size of each group from which to select customers from. A group size of 10 means to randomly select 1 customer out of every group of 10.
- Systematic Centered – For each stratum, specify the size of each group from which to select customers from. A group size of 10 means to select the 5th customer out of every group of 10.

LIStcount	x

LISTCOUNT (Optional) – Indicate the total number of lists to be generated. x must be an integer from 1 to 8. The default is 1, so if LISTCOUNT is not included, one list will be generated.

Sample B420 Stratification Control File:

VERSION	1, 2, 3, 4
GROUP_SIZE	1, 1, 1, 1
LISTCOUNT	1

Create the Stratification Control File with User Language

You can optionally compose the Stratification Control File with User Language statements. If you already have constructed your analysis control file, you can skip this section. The "User Language" allows for more customization, but keep in mind it is also more complex.

You build the Stratification Control File with User Language statements. A template for creating the Stratification Control File is shown in Figure 10-2. An example is provided in Figure 10-3.

A typical Stratification Control File is composed of a few basic elements, which are described below and illustrated in the template and example. You do not need to include selection statements (test statements that define eligible customers) in the Stratification Control File for a multidimensional design, because the new version of the Population Data File (from B220) is already limited to the target population.

Comments — optional notes that do not affect processing. You must enclose any comments between these symbols: /* and */. Do not split comments across lines, unless you begin and end each line with the slash/asterisk symbols. Do not enter the slash/asterisk symbols in columns 1 and 2 of the file; otherwise, you may put comments anywhere in the file.

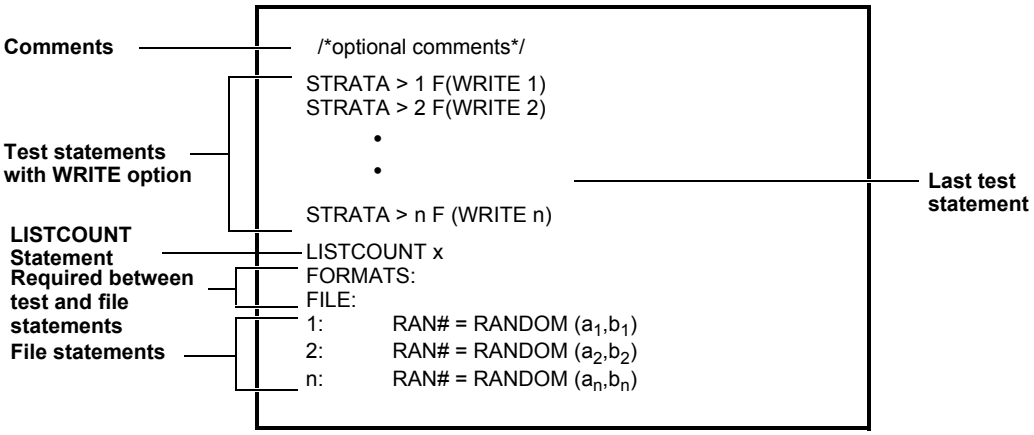


Figure 10-2 Stratification Control File "Template"

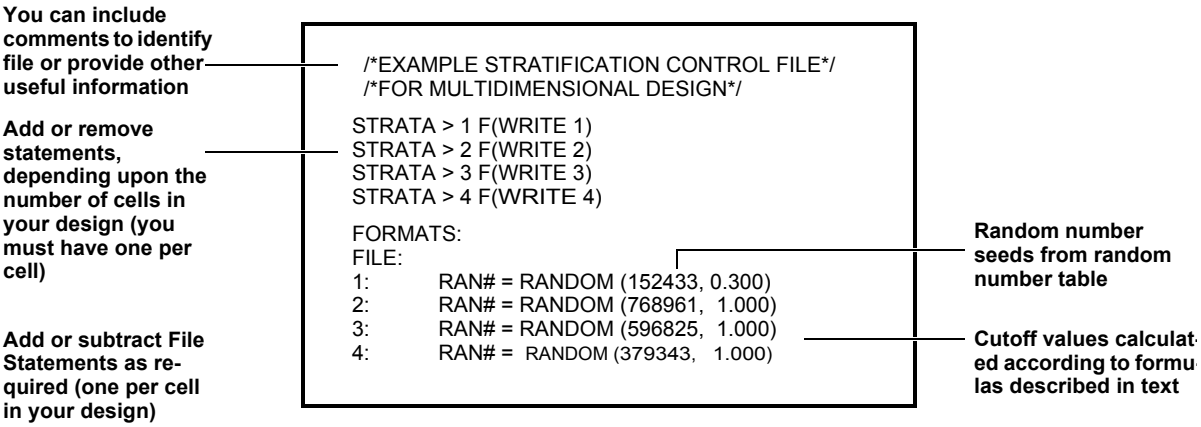


Figure 10-3 Example Stratification Control File

You may be able to create your own file by simply modifying this example. This example was created for a design with four cells.

Test statements with Write option — used to assign each customer a random number, according to its cell assignment.

Create a series of test statements using the following format. There must be one statement for each cell in your design.

STRATA > n F(WRITE n)

where **n** is a sequential number beginning with 1.

> **Note:** The statement uses the variable name "STRATA" because that is the field name in the PDF; but it is actually referring to the cell number as assigned by the Multidimensional Population Analysis Program.

Using the logic you construct, the program compares each customer's cell assignment number to the series of strata (e.g., cell) numbers. When the customer finally fails a test statement, it triggers the Write option. The Write option refers the program to the matching File Statement that assigns a random number.

The last test statement is typically:

STRATA > n F(WRITE n)

where n is the number representing the last cell in your design, (e.g., equals the total number of cells).

LISTCOUNT Statement— A statement that indicates the total number of lists to be generated. These lists can be treated as a primary list and x-1 alternate lists. The LISTCOUNT statement should be placed immediately before the DIM statement if there is one, or immediately before the FORMATS statement if the DIM statement is not present. Use this format to construct your LISTCOUNT Statement for sample selection:

LISTCOUNT x

where:

- **LISTCOUNT** indicates that you are creating more than one list.
- **x** indicates the total number of lists to be generated and must be an integer from 1 to 8. The default is 1, so if LISTCOUNT is not included, one list will be generated.

  Notes:

  - The lists generated by Single Dimension Selection B410 will be named TGB413.LST1 through TGB413.LSTx where x is the number of lists generated. The lists generated by Multi Dimension Selection B420 will be named TGB423.LST1 through TGB423.LSTx.
  - If a DIM statement and STATS statements are specified in the Report Control File then a Sampling Statistics File will be generated for each list. These will be named with suffixes .SSF1 through .SSFx where x is the number of lists generated.
  - The specifications coded in the Sample Selection control file should be based on the user's requirements for the primary list only. The Sample Selection program will use those specifications, combined with the value specified for LISTCOUNT, to select enough customers to fill all the lists requested, as long as enough eligible customers exist.
  - The lists will be filled in numerical order, so if the system runs out of eligible customers the higher numbered lists will be shorter than the lower numbered lists.

FORMATS:

FILE:

You must put these two keywords between the test statements and the file statements, just as shown in the examples.

File statements — a special kind of format statement that controls the content and organization of the output file (in this case, the Selection File). For each Write option in the test statements, you must have a matching file statement.

For Random Sampling, set up the file statement in the following format:

n: RAN# = RANDOM(a,b)

where:

- **n** is the number in the associated False clause, and therefore the cell number, and
- **a** is a random number seed that the program uses to initialize the random number process. **Specify the seed as a 6-digit, odd integer between 0 and 999999. Be sure to use a unique seed number for each statement.** You can obtain these seeds from a table of random numbers. ***Note:*** Each seed number generates the same sequence every time. This is useful should you need to re-create your process.
- **b** is a cutoff value that limits the number of records that the program must sort and report, saving processing time. (***Note:*** The use of a cutoff value does not mean systematic sampling. Instead, this process assigns all customers a random number and eliminates those with a low probability of selection.)

Specify the cutoff value as a real number between 0.0 and 1.0. If you specify a cutoff value of 1.0, every qualifying customer in the PDF will be selected for output in the Selection File.

How to calculate cutoff values: For large target populations (that is, greater than 30) you can calculate the cutoff value for a particular stratum using this formula: **c=n/N**

where:

- **n** is the required sample size for the stratum, including an allowance for alternates. You will typically use a value that is five times greater than the number of sample points required for the stratum, as calculated by the Sample Design Program. For example, if your design called for 20 sample points for the stratum, your value for **n** would be 100.
- **N** is the stratum population size.

For small samples, the cutoff value must be adjusted to ensure an adequate selection. In this case the appropriate formula is:

$$c = [\, n + 2 * SQRT\,(n - n^2 / N)\,] / N$$

where **n** and **N** are the same as described above. Typical cutoff values for small populations and sample sizes are presented in Table 10-1.

Table 10-1: Cutoff Values for Small Populations

	POPULATION SIZE (N)						
SAMPLE SIZE (n)	**50**	**100**	**250**	**500**	**1000**	**5000**	**100000**
10	0.3131	0.1600	0.0648	0.0325	0.0163	0.0033	0.0002
25	0.6414	0.3366	0.1379	0.0695	0.0340	0.0070	0.0003
60	1.0000	0.6000	0.2506	0.1260	0.0638	0.0128	0.0006
100		1.0000	0.4620	0.2358	0.1190	0.0240	0.0012
250			1.0000	0.5447	0.2774	0.0562	0.0028
500				1.0000	0.5316	0.1085	0.0054

Systematic Sampling selects the **R**th point from an ordered population, and every **k**th point thereafter. **k** is determined by dividing **N** by **n**, where **N** is the entire eligible population and **n** is the desired sample size. **R** is a random number between 1 and **k**, inclusive.

In effect, the population is divided into **n** groups, each containing **k** customers (with the last group possibly containing fewer than **k** customers). One customer from each group is chosen. For example, if the population consisted of 300 customers (**N**) and was divided into 30 groups, each group would consist of 10 customers, and selecting one customer from each group would yield a sampling population of 30 customers (**n**). When selecting one customer from each group, the Systematic Sampling method always selects the customer with the same relative position within each group. In other words, Systematic Sampling takes a random

number between 1 and **k**, where **k** is the size of each group, and selects the customer in each group whose position within the group is equal to that random number. So if, for example, the group size is 10 and the random number selected is 8, the 8th customer from each group is selected for inclusion in the sample population.

Unlike simple Random Sampling, Systematic Sampling forces an even distribution across the entire population from which sample points are drawn, providing a better representation of the population.

For Systematic Sampling, the File Statement is similar to that for Random Sampling, with this difference:

b is used to specify the size (**k**) of each of the groups into which the population is to be divided for selection.

The following File statements might be used for Systematic Sampling:

FILE:

1: STRATA = 1 RAN# = RANDOM (152433, 5)

2: STRATA = 2 RAN# = RANDOM (1768961, 10)

3: STRATA = 3 RAN# = RANDOM (596825, 7)

In this case, one customer would be selected from each group of 5 in the stratum 1 population, one from each group of 10 in stratum 2, and one from each group of 7 in stratum 3.

Centered Systematic Sampling is a refinement of the Systematic Sampling technique. In this case, the random number **R** is equal to **k / 2**. Unlike the Systematic Sampling method, Centered Systematic Sampling selects the middle customer of each group, rather than generating a random number to select a customer from each group.

The same File statements shown above for Systematic Sampling could be used for Centered Systematic Sampling, the difference being which customer would be selected from each group. For example, in stratum 1 the third customer from each group would be selected.

Note: For all three sampling methods, the random number is stored in its slot in the population record, even though the third method, Centered Systematic Sampling, does not use the random number. The population is later sorted by that random number, so random number seeds are still needed regardless of the sampling method used.

Step 1D: Create the Reporting Control File (.RCF)

During the Reporting Phase, the program creates a report of a user-defined number of selected customers, formatted to your specifications. The report typically includes (but is not limited to):

- Customer identifier or account number.
- Selection variables or other qualitative characteristics such as service company, rate class, and SIC code. If you are eventually doing systematic replacement, you will require user-defined variables for identification of like alternates (see **Step 6: Modify your Reporting Control File (.RCF)** on page 10-14 for more information about alternates).
- Additional data for installation or mailing purposes such as name, address, and telephone number.
- Cell number.

The program also calculates statistics for the selected sample, and outputs the statistics to the Sample Statistics File. These statistics are required for Sample Validation. Remember, if you plan to perform Sample Validation, you will need to give the Sample Statistics File a new dataset name.

Otherwise, this file will overwrite the Population Statistics File created by the Stratification Program.

Along with the Population Data File and the Record Definition File, the Reporting Phase requires a **Reporting Control File (.RCF)**.

You use the Reporting Control File to define the content and format of the selection report, including the number of candidates to be reported within each cell, and to output the required statistics. For the first program run, ***be sure to specify just the required number of sample points per cell***, so that the statistics created for validation will reflect the actual sample size. (You will have an opportunity to specify some number of alternates later in the process.)

Create the Reporting Control File

The Reporting Control File is created with the following commands:

USAge	usageVariable
SAMple_size	stratum1SampleSize, stratum2SampleSize … stratumNSampleSize
PRInt	printVariable1, printVariable2 … printVariableN
DELimiter	BLAnk
	COMma

USAge	usageVariable

USAGE – Defines the usage, or design variable. Specify the name of the usage variable you would like to use. For example:

USAGE JAN

SAMple_size	stratum1SampleSize, stratum2SampleSize … stratumNSampleSize

SAMPLE_SIZE – List the sample size for each stratum. Comma separated.

PRInt	printVariable1, printVariable2 … printVariableN

PRINT – Provide the list of fields to print for each selected customer in the final selection report. Comma separated.

DELIMITER	BLAnk

DELIMITER – (optional) controls the formatting of the output list file(s). **BLAnk** indicates to produce a space delimited output file. This is the default. **COMma** produces a comma separated output list file.

Sample RCF File

USAGE	JAN
SAMPLE_SIZE	85, 15, 36, 34
PRINT	STRATA, CUSTID, NAME, ADDRESS

You can optionally compose the Reporting Control File with User Language statements. If you already have constructed your Reporting Control File, you can skip this section. The "User Language" allows for more customization, but keep in mind it is also more complex.

You create the Reporting Control File with User Language statements. An example Reporting Control File is shown in Figure 10-5. To get started, you may be able to simply modify the example Control File shown here. It is likely however, that you will wish to create a more detailed report tailored to your specific needs. The User Language is very flexible and will enable you to output any data in your Population Data File, formatted according to your own design. **To produce more complex or customized reports, you may find it necessary to refer to the User Language instructions provided in Appendix A**: **The User Language.**

Important Note about DIM Statement: To generate the Sample Statistics File required for Sample Validation, you must include the DIM Statement and STATS option, as shown in Figure 10-4 and Figure 10-5. To create your DIM Statement, follow the instructions provided in Step 1C in **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**.

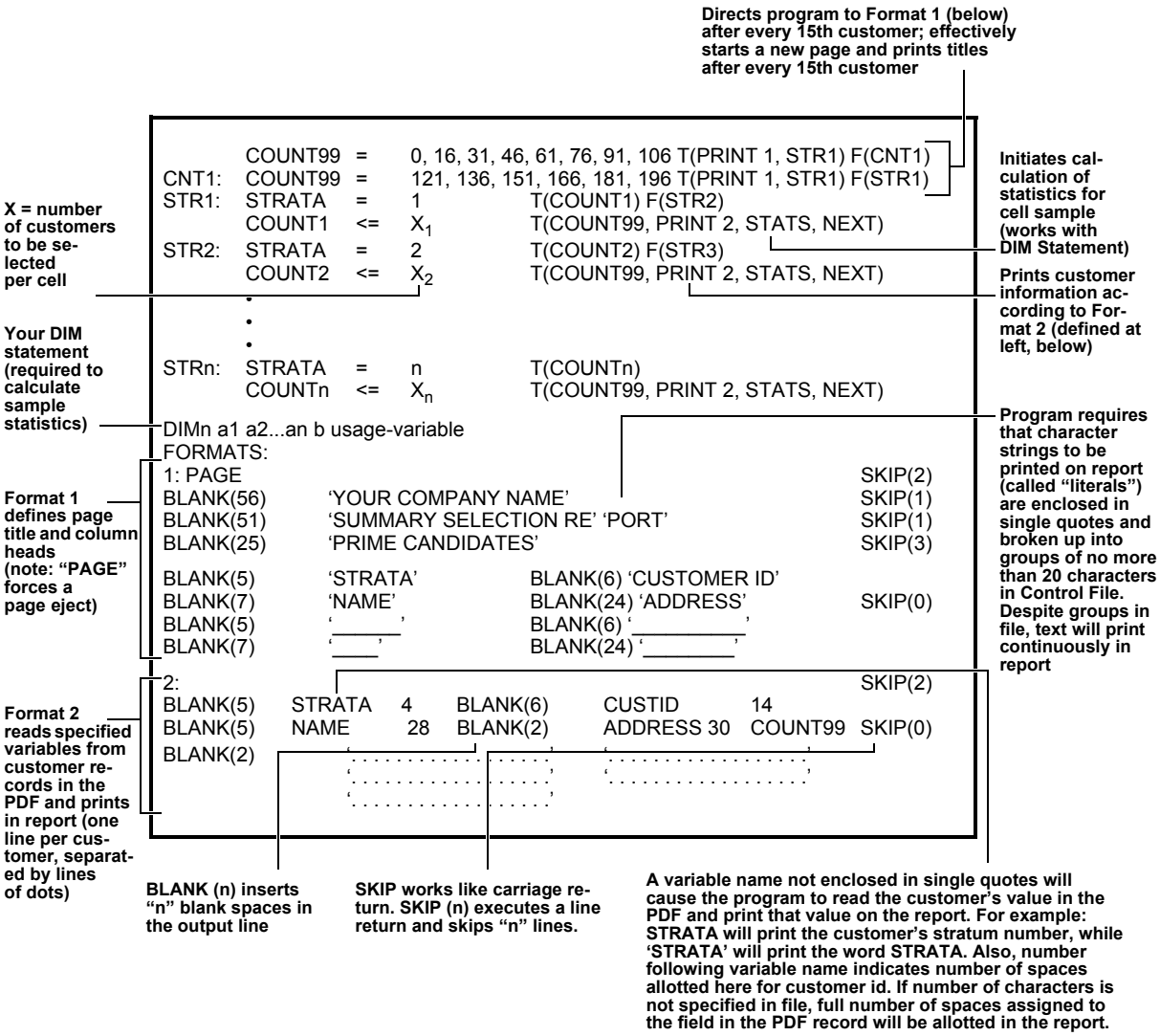


Figure 10-4 Reporting Control File "Template"

This template produces a Sample Statistics File and a report with 15 customers per page. It assumes that variable names in PDF for customer ID, name, and address are CUSTID, NAME, and ADDRESS, respectively.

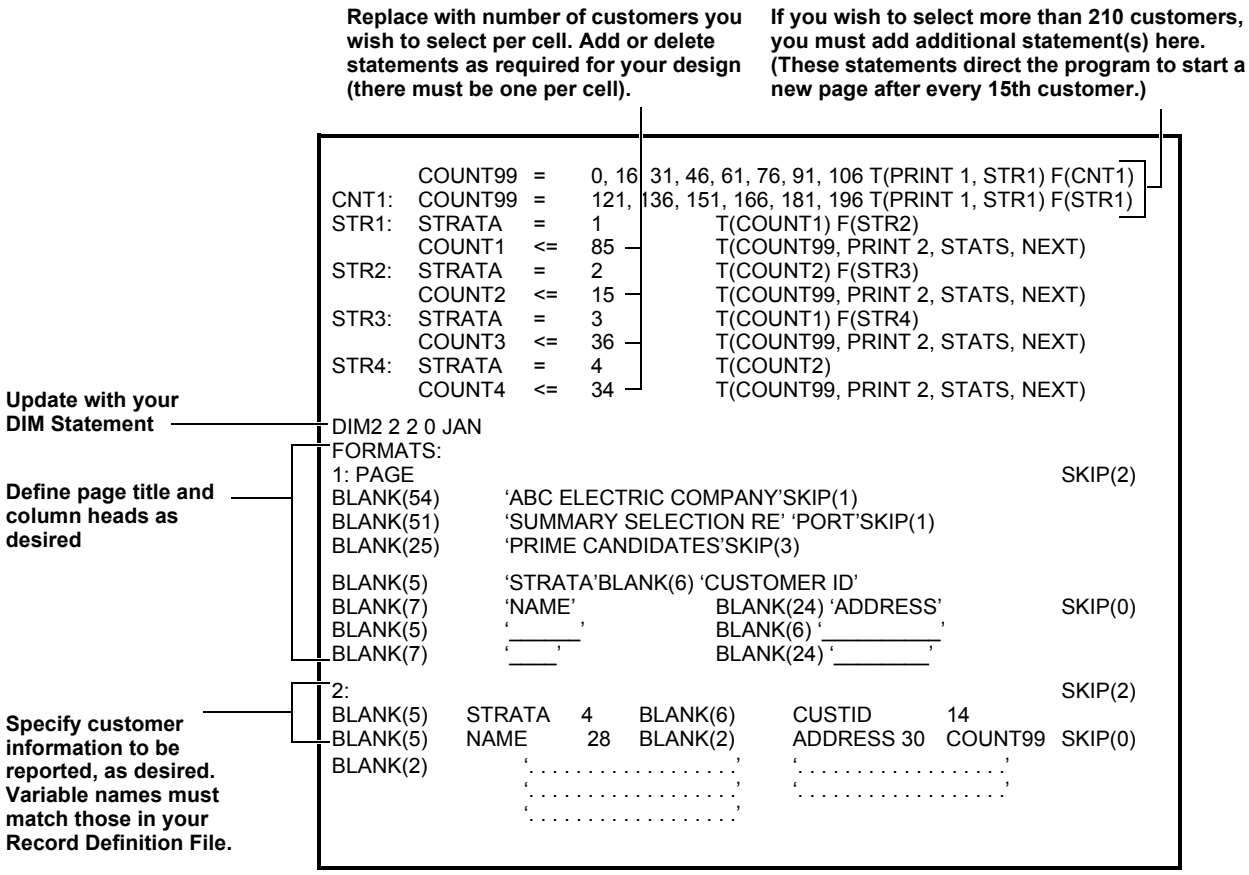


Figure 10-5 Example Reporting Control File

This example was used to produce a Sample Statistics File based on January usage. It illustrates how you can use the template to create your own file and report.

Step 2: Submit the Sample Selection Procedure

Use the B420 **Submit Panel**.

The first page of the B420 **Submit Panel** contains the following fields:

- Control File - the Stratification Control File you created in Step 1C (e.g., TGB42A.CTL)
- Report Control File - the .RCF Reporting Control File you created in Step 1D
- Record Definition File - the .DEF file created from B110
- Scored Population Data File - the .PDS file from B220
- Sample Statistics File - optionally specify a name for the .SSF file output
- Selection Method: select RANDOM, SYSTEMATIC, or CENTERED.

Fill in or select values for these fields, then click the **Submit** button to submit the Sample Selection job.

Step 3: Check Output

Each program in the Sample Selection Procedure produces its own set of outputs.

Stratification Program

- **Population Analysis Execution Log (TGB220-01)** — contains a copy of the Stratification Control File and any processing messages.
- **Selection File (SORTED.PDS)** — sorted copy of the PDS, with the reserved field RAN# now filled in. Only the number of customers you requested in the Stratification Control File (via the cutoff value) are output to the Selection File.

Reporting Program

- **Population Analysis Execution Log (TGB220-01)** — lists the Reporting Control File, any processing messages, and the desired customer selection report.
- **Sample Statistics File (.SSF)** — Strata means and standard deviations for the sample. You will use this file as input to the Sample Validation Program.
- **Table of Cell Entries Report** — printed version of Sample Statistics File in a slightly modified format.

Step 4: Rerun Sample Selection to Calculate Statistics

You can rerun sample selection to calculate sample statistics for each additional usage variable in your design. Only two modifications are required between program runs:

- You must identify a different usage variable in the Reporting Control File's USAGE (or DIM Statement if you are using the User Language).
- You must assign a new dataset name to the Sample Statistics File to avoid overwriting statistics calculated in earlier runs.

Optional: Perform the following additional steps if you wish to draw alternates:

Step 5: Validate your Sample Selection using the Sample Validation Program

Before drawing alternate sample points, it is important that you first validate your initial sample using the statistics produced by the Sample Selection Procedure. Follow the instructions provided in **Chapter 11**: **Validating the Sample**. You must validate the sample for each usage variable.

- *If the sample passes validation for all usage variables in your design*, go to Step 6 below.
- *If the sample fails validation*, replace the version number (random number seeds if you are using User Language) in your Stratification Control File (TGB42A.CTL) with a completely new set, and rerun the Sample Selection Procedure. Repeat this process until you arrive at a satisfactory sample selection, then proceed to Step 6.

Step 6: Modify your Reporting Control File (.RCF)

You can modify your reporting control file (.RCF) to specify an allowance for alternates. For your final sample selection, you will typically request the number of sample points specified by the sample design program *plus* an allowance for alternates. Typically, not all customers selected will be willing or able to participate in the sample, so you need to add some allowance that ensures getting the required number of points in the field and collected. The pool of alternates will be selected in sort order, based on the random number assignment.

In your Reporting Control File, update the number of customers to be selected per cell. The numbers you specify will depend upon your criteria for selecting alternates:

- *If you intend to replace primary sample points with the next random selection,* we recommend that you select two to three times the number of customers required to meet the desired accuracy.
- If you must select alternates with like characteristics (that is, systematic replacement), you will need a large pool from which to draw. We recommend that you select five times the number of required sample points. When defining the Reporting Control File, keep in mind also that you must include the pertinent selection variables in your report, so that you can identify like customers. For example, a criteria for drawing alternates might be that they share the same

meter reading route, in which case you would have to identify each customer's route on the report.

Note: If you request more customers than have been assigned to a cell, the program will simply select all members of that cell. The program will not issue an error message.

Step 7: Resubmit the Sample Selection Procedure (B420) and Check Output

You have now completed the sample design, selection, and validation process. The result should be a sample that accurately mirrors the targeted population and meets all design criteria.

Chapter 11

Validating the Sample

The **Multidimensional Sample Validation Program** enables you to check the results of your sample selection *for a simple random, single dimensional stratified, or multidimensional stratified design*. The program evaluates how well the selected sample mirrors the population in terms of the design variable.

Using a Z-test, the program compares the mean of the design variable for the sample to that of the population, strata by strata or cell by cell. Taking standard deviation into account, the program evaluates whether or not the two means are significantly different within each stratum or cell. If there is a significant difference, the program gives you a warning message.

Secondarily, the program also computes the expected accuracy of the sample within each stratum or cell. These individual accuracies are less than the total sample accuracy that was used as a design criterion. However, it provides a relative measure for comparative purposes during the design process.

Here is a summary of the steps to follow for the Sample Validation Program. If you are validating a multidimensional design, you must run the Sample Validation Program *for each usage variable in your design*. The only difference between the program runs is the statistics files used for input: each time you run the program to validate the sample for a given usage variable, you must supply the Population Statistics File and Sample Statistics File computed for that variable.

SUMMARY
VALIDATING THE SAMPLE SELECTION
USING THE SAMPLE VALIDATION PROGRAM (B520)

1. Verify or create the required input files:
   - Sample Validation Environment File (.RAF) — specify desired level of precision for your sample; use relative Accuracy File created by B320 for a multidimensional design, or create file.
   - Sample Statistics File (.SSF) —— use the .SSF file from Step 3 in B410 for a single dimensional design or B420 for a multidimensional.
   - Population Statistics File (TGB52C) — use the .PSF from Step 1 in B410 for a single dimensional design; STEP3.PSF from B220 for a multidimensional design.
2. Submit the job (Multidimensional Sample Validation Program — B520).
3. Check output:
   - Multidimensional Sample Validation Environment Report (TGB520-01)
   - Multidimensional Sample Validation — Population File Report (TGB520-02)
   - Multidimensional Sample Validation — Sample File Report (TGB520-03)
   - Multidimensional Validation Report (TGB520-04) — includes Relative Accuracy Report and Sample Validation Report.

If you have a multidimensional design with more than one usage variable:

4. Repeat the steps 1 - 3 for each additional usage variable in your design (using a different set of Population Statistics and Sample Statistics Files).

Step 1A: Create or Verify the Sample Validation Environment File (TGB52B.RAF)

The Sample Environment File specifies the desired level of precision for your sample and optional report titles. It **must** contain the same precision you specified for the Design Command in the Sample Design Environment File (**Chapter 5**: **Stratifying the Population and Determining Sample Size for a Single Dimension** for a single dimensional design, **Chapter 8**: **Determining Sample Size for a Multidimensional Sample Design** for a multidimensional). This file was an output from Multidimensional Sample Design (B320). If you are validating a single dimensional design, you can create the file manually. Up to three Group comments may be supplied with titles that will be displayed at the top of each report. Each title may be up to 76 characters in length. The format of the commands is:

ALPha [5.00 | <u>10.00</u>]

GROup *title*

Step 1B: Verify the Sample Statistics File (.SSF)

This file was generated by the Sample Selection Program, based on the STATS option you incorporated in the Reporting Control File (see **Chapter 9**: **Selecting the Sample for a Single Dimensional Design** for a single-dimensional design, **Chapter 10**: **Selecting the Sample for a Multidimensional Design** for a multidimensional). Use the .SSF from Step 3 in B410 or B420. The file contains statistics about the selected sample, stratum by stratum or cell by cell. It is organized as follows:

Column 1	*Cell or strata number:* same as that written out to the Population Data File in the STRATA field.
Column 2	*Cell/stratum population:* a count of the number of customers selected to represent that particular cell or stratum.
Column 3	*Cell/stratum mean:* mean of the design usage variable for all selected customers in the cell or stratum.
Column 4	*Standard deviation in cell or stratum:* standard deviation for the design usage variable for all selected customers in the cell or stratum.
Column 5	Number of observations in the Population Data File.

Verify that the Sample Statistics File contains one line of information for each cell or stratum in your sample design.

Step 1C: Verify the Population Statistics File (.PSF)

This file should have been generated by the Sample Selection Program for a single-dimensional design (see **Chapter 9**: **Selecting the Sample for a Single Dimensional Design**) or the Multidimensional Population Analysis Program for a multidimensional design (see **Chapter 8**: **Determining Sample Size for a Multidimensional Sample Design**). Use the .PSF file from Step 1 in B410 for a single dimensional design, and the .PSF file from B220 for a multidimensional design. An example is shown in Figure 11-1.

CELL(1—1)	1	173	484.21	406.52	173
CELL(1—2)	2	2	3699.00	2018.08	2
CELL(2—1)	3	672	4517.15	458.02	672
CELL(2—2)	4	52	4949.02	4497.65	52

Figure 11-1 Example of a Population Statistics File

The file contains statistics for all customers in the PDF that were assigned to each particular cell or stratum. The file is organized as follows:

Column 1 — *Dimension indices* indicates the strata number in each dimension that defines the cell.

Column 2 — *Cell or strata number:* same as that written out to the Population Data File in the STRATA field.

Column 3 — *Cell/stratum population:* a count of the total number of customers in the PDF that were assigned to that particular cell or stratum.

Column 4 — *Cell/stratum mean:* mean of the design usage variable for all customers in the cell or stratum.

Column 5 — *Standard deviation in cell or stratum:* standard deviation for the design usage variable for all customers in the cell or stratum.

Column 6 — Number of observations in the Population Data File.

Verify that the Population Statistics File contains one line of information for each cell or stratum in the sample design.

> **Note:** Even if you used statistics from prior load research data to size your sample, be sure to use the Population Statistics File that was calculated for your target population. Use the version that was created by the Sample Selection Program (for a single-dimensional design) or Population Analysis (for a multidimensional), rather than the statistics created with one of the Oracle Utilities Analysis programs.

Also, if you updated the Population Statistics File to specify 100% sampling for the Multidimensional Sample Design Program, be sure to remove the "100%" entries before using the file here.

Step 2: Submit the Job (B520)

Submit the Job (B520).

Step 3: Check Output

The Sample Validation Program produces four reports:

- **Sample Validation Environment Report (TGB520-01)** — indicates the level of significance desired for the relative accuracy computation and sample validation, as specified in the Sample Validation Environment File.
- **Validation Population Report (TGB520-02)** — provides a detailed breakdown of the population statistics within each cell or strata, as provided in the Population Statistics File.
- **Validation Sample Report (TGB520-03)** — provides a detailed breakdown of the sample statistics by cell or strata, as provided in the Sample Statistics File.
- **Sample Validation Report (TGB520-04)** — consists of two tables. The first shows the expected relative accuracies by cell or strata; the second indicates the validity of the selected sample based upon KWH.

When evaluating the Relative Accuracy Table, the critical values are the totals for the sample, not those for individual cells or strata. To evaluate the validity of the sample, compare the value for the relative accuracy of the total (e.g., sample) to the desired accuracy.

> **Note: If your sample selection fails validation**, you must redraw the sample using a completely different set of random number seeds in your Stratification Control File (see Step 4 in **Chapter 9**: **Selecting the Sample for a Single Dimensional Design** for a single-dimensional design, Step 5 in **Chapter 10**: **Selecting the Sample for a Multidimensional Design** for a multidimensional).

Step 4: Repeat the Preceding Steps

If you have a multidimensional design with more than one usage variable:

Repeat the preceding steps for each additional usage variable in your design. You must perform Sample Validation for each usage variable in your design. The only difference between program runs is the set of Population Statistics and Sample Statistics files you use as input. In other words, each time you run the program to validate the sample for a given usage variable, you must supply the Population Statistics File and Sample Statistics File computed for that variable.

Appendix A

The User Language

The User Language is a collection of statements that control processing of the Population Data File (PDF). It is a very flexible and powerful language that you use throughout the sampling process for a variety of purposes: to select members of the PDF for inclusion in the sampling process, to assign customers to strata or cells, to calculate stratum or cell statistics, and more. You can also apply the User Language to create your own customized reports of any information in the PDF, to output PDF data to files for use by other programs, to count the occurrences of a specified item or condition in the PDF, or to modify contents of PDF fields.

Throughout this guide, elements of the User Language have been explained in the context of specific tasks (creating an Analysis Control File to generate a frequency distribution table, for example). In this appendix, we examine each of the basic "building blocks" of the User Language in detail, so that you will have a thorough command of the language and can apply it to your own unique needs.

How It Works

Essentially, the User Language enables you to identify customer records by desired characteristics, and to specify the actions the program will take once those records have been identified.

You place the User Language Statements in a Control File that becomes input to one of the Sampling Programs (Population Analysis or Sample Selection). Your Control File acts as a processing template for each customer record in the PDF. That is, the program evaluates each PDF record starting with the first "test statement" in your file. Based on the outcome of an individual test (whether or not the customer record meets the criteria), the User Language may:

- Branch forward to another test statement
- Update a counter variable (count the occurrence of a specified item or condition)
- Compute cell- or stratum-specific statistics
- Print selected PDF fields on a Report File
- Update and write the PDF record to a Selection File
- Restart testing with the next PDF record
- Terminate all testing
- A combination of the above.

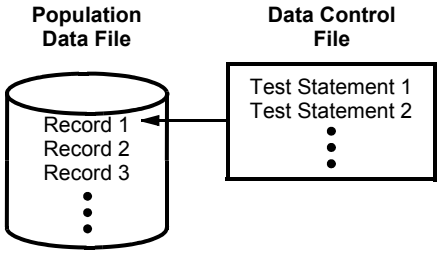


Figure A-1 User Language

Use the User Language to construct a list of true/false test statements that establish your criteria for desired records. The Sampling Program applies each test statement to the first record to determine whether or not it matches your criteria. After testing the first record, the program applies the set of test statements to the second record, then the third, and so on until the program has examined the entire database file, or it encounters a STOP Command.

Major Statement Types

As illustrated in the diagrams below, the User Language consists of five major statement types that you can use to build your Control File. Whether or not you use all five types in a single Control File will depend upon your objectives.

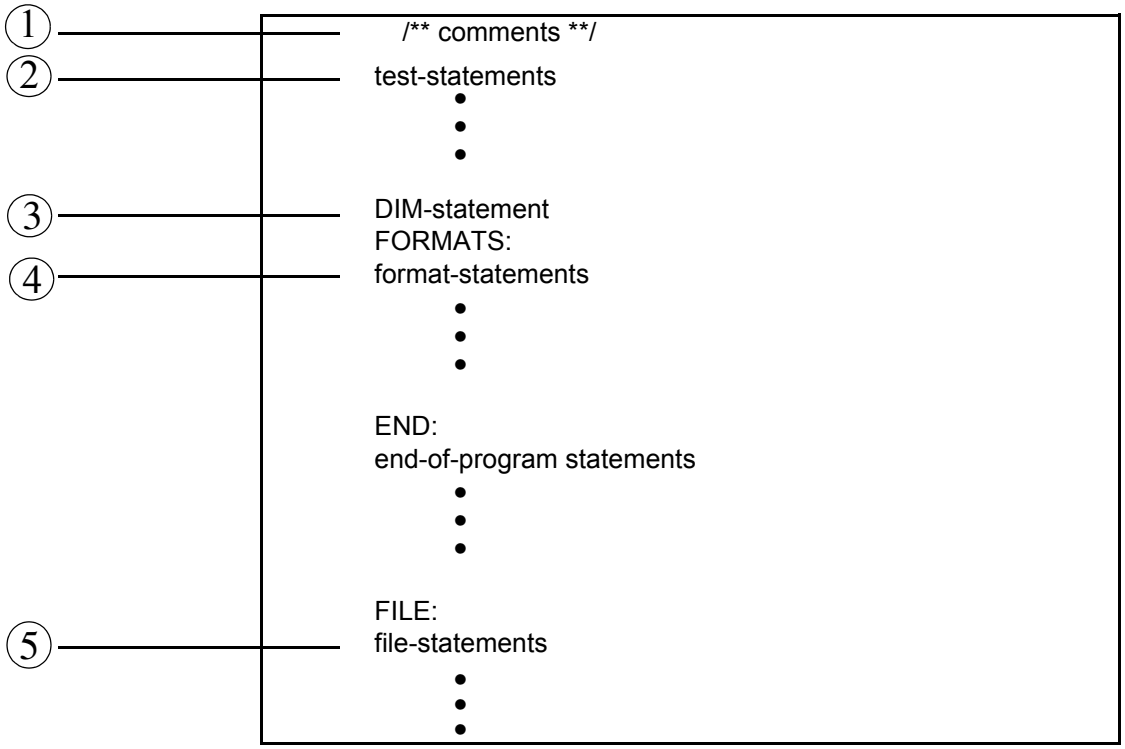


Figure A-2 Control File Format

Each of the numbered items is referenced in the text.

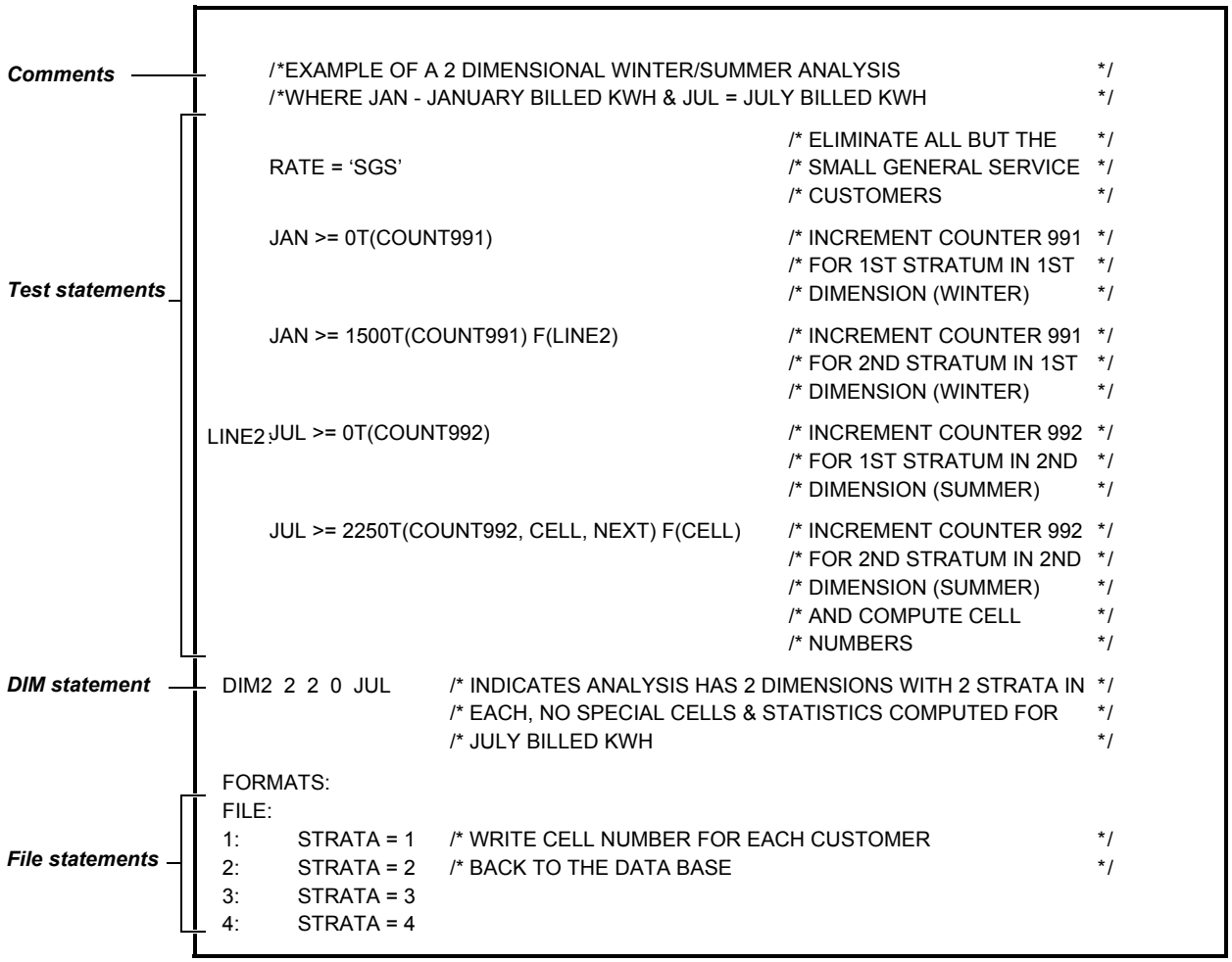


Figure A-3 Example Control File Illustrates many User Language Elements

In your Control File, test statements **must** precede format and file statements. Comments can be placed anywhere in the file. Let's look briefly at each statement type:

1. **Comments** — optional notes that do not affect processing. You can use them to record the purpose of the file, for example.
2. **Test Statement** — compares the contents of a selected field in a Population Data File record (called the "variable") to your criteria (called the "test value"), and indicates what action should be taken depending upon the results of the comparison. Variables are defined in your Population Record Definition File; e.g., variable name, data type, length, etc.
3. **DIM Statement** — specifies the dimensions of a sample. Required only when you wish to compute cell or strata specific statistics, or assign customers to cells or strata. The DIM Statement is used in conjunction with the CELL or STATS options in a test statement.
4. **Format Statement** — specifies how the results should be organized in a report or output file. You can specify what information you want printed (including data from any fields in the selected records, as well as titles and notes) and how you want the information organized in the file or on the page. Format statements are typically activated by a PRINT option in the test statements. A special type of format statement, called the "end of program statement,", is initiated when the PDF processing has been completed. They are generally used to print summary information.
5. **File Statement** — specifies how the results should be organized in an output file called the "Selection File." A special type of file statement is used to replace the contents of PDF fields

with new values (such as STRATA number, for example.) File statements are activated by a WRITE option in the test statements.

Counter variables are also important. They are special elements that you can use in different ways throughout the file — for example, to count the number of PDF records with specified characteristics, or to determine strata assignments based on the number of times a given customer triggers a counter.

Each of the major statement types has many options and can be complex, so we will examine each of them more closely in the following sections.

> **Note: Blank lines** may be used freely throughout the file to improve readability. They are ignored during processing.

Comments

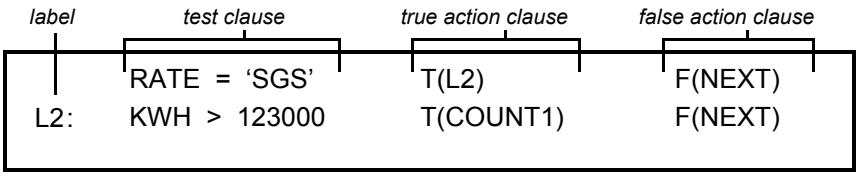
Comments are optional notes that you can use to record the purpose of a file or other information. You **must** enclose your comments in these symbols: /* and */. Do not split comments across lines, unless you begin and end each line with the slash/asterisk symbols. Do **not** enter the slash/asterisk symbols in columns 1 and 2 of your file as it will cause the program to stop.

Test Statements

Test statements enable you to select or analyze a customer population by desired characteristics, and to specify what action should be taken depending upon how each customer record compares to the specified criteria.

Each test statement consists of three major elements: a test clause and two action clauses. The **test clause** compares the contents of a selected field in a Population Data File record (the "variable") to one or more test values. The comparison is either true or false. Depending on the result, the test statement activates either the **true action clause** or the **false action clause**. In addition, test statements may be identified by an optional label for forward branching.

Consider the following example:



The first test statement says that for each record in the Population Data File, look at the customer's value in the field named RATE. If it is true that it matches SGS, then go to the test statement labelled L2 and continue testing that customer record. If it is false, skip that customer and test the next record in the PDF.

When creating your own test statements, you may use blanks or commas to delineate the test statement options. You can input up to 200 test statements in a single Control File, but only one test statement per line.

There are many options when creating a test statement, particularly in the action clauses. The complete test statement format is illustrated below:

[test-label:] [test-clause] [true-action-clause] [false-action-clause]

Test clause is in the format:

variable [relation] test value[[relation] test value] . . .

True-action-clause is in the format:

T ([**PRINT** *n*] [**WRITE** *n*] [,**COUNT** *m*] [,*label* | ,**STOP** | ,**NEXT** | , continue])

False-action-clause is in the format:

F ([**PRINT** *n*] [**WRITE** *n*] [,**COUNT** *m*] [,*label* | ,**STOP** | ,**NEXT**])

To better understand how you can construct your own test statements, let's look at each element and option in detail.

- **Test label** — Identifies the test statement for reference by other statements. Labels are optional; you need to preface a statement with a label *only* if it will be the target of other statements. A label can be any word or code of your own choosing, as long as it is no more than 8 alphanumeric characters long, and the first character is alphabetic. Input a colon (:) after the label, but do **not** insert a blank between the label and the colon. Here are three sample labels:

  L1:

  LINE:

  STR1:

Test Clauses

variable [relation] test value[[relation] test value] . . .

- **Variable** — The name of PDF record field being examined, as defined in your Record Definition File. ***Note:*** You must spell the name as it appears in your Record Definition File. For example, if you were using the file shown in Figure 3-1, eligible names would be CUSTID, RATE, ROUTE, and so forth.
- **Relation** — How the record value should compare to the test-value(s). You can use any of the following relations. "Equal to" is the default.

=	variable equal to test-value
>	greater than
<	less than
#	not equal to
>=	greater than or equal to
<=	less than or equal to

- **Test Value** — Criteria for the evaluation. It can be a constant or character string which may exist in the PDF, or it can be a second variable. For example, consider the following test clauses:

  KWH > 123000

  REGION = 'NORTH'

  JANKW < MAXKW

The first two example test clauses compare a variable to a constant and a character string, respectively. The third example compares a variable to a second variable; e.g., compares the customer's value for its January bill to its value for contract maximum KW. (***Note:*** The User

Language automatically consults your Record Definition File to determine whether the test value is a constant or another PDF field. It also looks up the data type of the variable prior to the comparison and automatically converts the data type of the test value to match.)

IMPORTANT! *If the variable in the test clause was defined as character type data in your Record Definition File and your test value is a character string, you must enclose the test value in single quote marks.* For example, RATE = 'SGS'. (If you refer back to our sample Record Definition File in Figure 3-1, you'll see that RATE was specified as CHAR.)

About comparisons with numeric data: You must take special care when specifying test values for comparison with numeric data *if* the variable in your PDF has been defined as character type data. This is because character type data is stored character by character, left justified, and padded with blanks. During the comparison to the test value, data in the PDF field will be evaluated left to right. Therefore, when specifying your test value, you must enclose it in single quotes *and* match the left or right justification in the PDF. (If the numeric field was defined as packed, integer, or real type data, you do not need quotes or attention to justification when building the test clause.) Examples of test clauses applied to numeric data that have been specified as character type data, CHAR(6), in the PDF:

JAN > ' 0'

JAN > ' 1300'

JAN > ' 0000'

About multiple comparisons: You can include multiple comparisons within a single test clause. In such cases, a logical OR is assumed (test-value1 or test-value2, or...etc.). The comparison continues until the condition is satisfied or the expression is exhausted. Examples of valid test clauses with multiple comparisons:

RATE = 'A1','A2','A3'*RATE is equal to A1, A2, or A3*

SIC = '5100' '5200'SIC is equal to 5100 or 5200

About testing for a range of values: If you wish to select customers based on a range of values, you must put the values in separate Test Statements. Examples:

STRATA >= 2 STRATA <= 4	Together, these two statements would select all customers in strata 2 through 4, inclusive.
JAN > 100 JAN < 500	These two statements would select all customers whose January usage was between 100 and 500 kWh exclusive.
JAN > 100 < 500	This single statement would **not** work, since all values satisfy the condition "greater than 100 OR less than 500".

Action Clauses

Each time a PDF record is compared to a test clause, the result is either a true or false evaluation. (If you omit the test clause, the evaluation defaults to "true".) You use the true action clause to specify what actions the program will take if the evaluation is true; use the false action clause to specify what actions the program will take if the evaluation is false.

The action clause formats are illustrated below. Use the prefixes "T" and "F" to identify the action clauses as true and false, respectively. The parentheses are required around the action options, and **no** intervening blanks may appear between the prefix and the left parenthesis. Other than the prefixes, the two action clauses differ only in their defaults (underlined).

True action clause

T ([**PRINT** *n*] [**WRITE** *n*] [,**COUNT** *m*] [,*label* | ,**STOP** | ,**NEXT** | , continue])

False action clause

F ([**PRINT** *n*] [**WRITE** *n*] [,**COUNT** *m*] [,*label* | ,**STOP** | ,**NEXT**])

You can include the PRINT n, WRITE n, COUNTm, or STATS options with any one of the other options in the same clause. However, the "label", "STOP", "NEXT", and "continue" options are mutually exclusive.

Logically, each test statement must have one true clause and one false clause. However, in many cases you will not actually need to input both, since you can often rely on the defaults. *If the true clause is omitted,* the default action is to continue processing with the following test statement. *If the false clause is omitted,* the default action is F(NEXT); i.e., to restart the test sequence with the next PDF record. If there are no more test statements, the implied action is also to restart the test sequence with the next PDF record.

The following describes each of the action options in detail. Some options (PRINT n, WRITE n, and COUNTm) are general purpose tools you can use in a variety of ways to create your own reports, output data to a file, modify PDF data, or analyze PDF data; while others (CELL and STATS) have more specific and limited applications in the Sampling process.

- **PRINT n** — directs the program to write data from a PDF record, or execute other format features (pagination, titles, notes, etc.), according to the format you specify in format statement "n:".

  Input the word "PRINT" followed by a blank and the required parameter "n", an integer between 1 and 99 inclusive that refers to a labelled format statement in your Control File. You can have no more than one PRINT option per action clause.

  Here is an example PRINT option and its corresponding format statement:

RATE = 'A1' T(PRINT 1) FORMATS:	*This set of User Language Statements will print a list of all customer ids in rate code A1.*
1: CUSTIDSKIP(0)	*Note how the PRINT 1 option refers to the format statement labelled "1:".*

- **WRITE n** — directs the program to assign new values to PDF fields and to write the resulting updated PDF record to an output file, according to instructions you provide in file statement "n:".

  Input the word "WRITE" followed by the required parameter "n", an integer between 1 and 99 inclusive that refers to a labelled file statement in your Control File. You can have no more than one WRITE option per action clause.

Here are some example WRITE options and their corresponding file statements:

```
STRATA > 1  F(WRITE 1)

STRATA > 2  F(WRITE 2)

STRATA > 3  F(WRITE 3)

FORMATS:
FILE:

1:  RAN# = RANDOM(152433, 0.300)

2:  RAN# = RANDOM(768961, 1.000)

3:  RAN# = RANDOM(596825, 1.000)
```

This set of User Language statements will create a new file in which each customer's random number field (RAN#) will be updated with a random number according to that customer's stratum (value in STRATA field). Note how the WRITE n options refer to the file statements labelled with a corresponding number "n".

- **COUNTm** — directs the program to count the number of occurrences of the condition specified in the test statement. Each time a PDF record is tested and matches the condition, the m^{th} counter variable is incremented by 1.

  Input the word "COUNT" followed by the required parameter "m", an integer between 1 and 999 inclusive. Do not insert a blank between COUNT and m. You can have no more than one counter per action clause.

  Here is an example use of counters:

  ```
  RATE = 'A1'  T(COUNT1)

  FORMATS:

  END: 'NUMBER OF CUSTOMERS ' 'IN RATE A1 IS ' COUNT1
  ```

  This set of User Language Statements will count the number of customers in rate code A1 and will output the resulting count in a report. If the number of customers in rate code were 44, the resulting line in the report would be: NUMBER OF CUSTOMERS IN RATE A1 IS 44

  Counters have many other applications in the Control File besides action options. They are explained later in this appendix under **Counter Variables** on page A-15.

  Counters 991 through 999 are reserved for a special application: they are used with the CELL option to assign customers to the appropriate cell or stratum and to compute population statistics. These counters are unique because they automatically reset to zero before a new PDF record is processed. They are required in the Population Analysis Control File (TGB22A) for Multidimensional Population Analysis (PROC STARB200 or EXEC B200 — **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**), and in the Stratification Control File (TGB22A) for Single Dimensional Sample Selection (PROC STARB410 or EXEC B200 — **Chapter 9**: **Selecting the Sample for a Single Dimensional Design**). See Figure 6-2, Figure 6-4, Figure 9-2, and Figure 9-3 for examples. See *"Step 1C"* in **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics** for a more complete description of how these counters work.

  Note: COUNT998 and COUNT999 are reserved for use with special cells.

- **CELL** — the CELL option is used to assign customers to the appropriate cell or stratum based on the values in the counter variables COUNT991 - 999, and to direct the program to compute the population statistics (number of customers, mean and standard deviation) for each cell or stratum. The CELL option is employed in the same files identified above for "Counters 991 - 999."

  When you use the CELL option, your Control File **must** also include counter variables COUNT991 - COUNT999 (one for each dimension in your design) and a DIM Statement (the DIM Statement indicates the number of dimensions, the number of strata within each

dimension, the number of special cells, and the usage variable for which the statistics are to be calculated). You **must** specify the CELL option in test clauses so that it is invoked for each PDF record in the population exactly once. This is usually done by inserting the CELL option as *both* the True and False action in the last test statement in the set of statements designed for cell/stratum assignment. See Figure 6-2, Figure 6-4, Figure 9-2, and Figure 9-3. For a description of how the Cell option works in conjunction with the counters and DIM Statement, see *"Step 1C"* in **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**.

- **STATS** — the STATS option is used to compute sample statistics for each stratum or cell in the design (these statistics are required for Validation). This option refers to the value in the PDF STRATA field to identify each customer's cell or stratum assignment, and then calculates the sample mean and standard deviation in each cell or stratum for the usage variable identified in the DIM Statement.

  When you use the STATS option, you **must** supply a DIM Statement in the Control File. The stratum (or cell) numbers must have already been written to the STRATA field in the PDF.

  The STATS option is used specifically in the Reporting Control File (TGB22A) for Sample Selection (PROC STARB410 and STARB420; EXEC B410 and B420 — **Chapter 9**: **Selecting the Sample for a Single Dimensional Design** and **Chapter 10**: **Selecting the Sample for a Multidimensional Design**). See Figure 9-4, Figure 9-5, Figure 10-4, and Figure 10-5 for examples.

The following action clause options are used to direct processing after the program has evaluated a PDF record according to the test clause. You must include one of them in each action clause, or rely on the default.

- **Label** — directs the program to compare the current PDF record with the test statement prefixed by "label:". As explained earlier, the label can be any word or code of your choosing, up to 8 characters long; the first character **must** be alphabetic. The labelled test statement must occur below the clause in the Control File; in other words, only forward branches are allowed. An example:

  ```
  RATE        =    'SGS'       T(L2)
  L2:  KWH    >    123000      T(COUNT1)
  ```

- **STOP** — executes any end-of-program format statements in the Control File and then halts the program. It is used to avoid processing the remaining PDF records once the testing condition has been satisfied. For example, this option would avoid scanning the entire database if you were searching for just one particular record. An example:

  ```
  CUSTID    =    'B176509'                T(STOP)
  ```

- **NEXT** — terminates the evaluation of the current PDF record and restarts the testing sequence with the next PDF record. This action is the default for all false clauses and the last true clause in the Control File. For true clauses, input the word NEXT; for false clauses, no input is required.

  An example:

  ```
  KWH <= 100                          T(COUNT1,NEXT) F(LINE2)
  LINE2:  KWH <= 200      T(COUNT2,NEXT) F(COUNT3,NEXT)
  ```

- **continue** — directs the program to compare the current PDF record with the next test statement. This action is possible only for true clauses and is the default for all true clauses except the last one in the file. (Because it is the default, no entry is required. Do **not** input the word "continue".) To continue when a false clauses is executed, you must put in the label of the next statement to be applied.

More About Action Clause Defaults: Remember, if the true clause is omitted, the default action is to continue processing with the following test statement. If there are no more test statements, the implied action is T(NEXT); to restart the testing sequence with the next PDF record. If the false clause is omitted, the default action option is F(NEXT). A simple example illustrates these default actions:

```
KW >= 10
KW > 50          T(NEXT)          F(COUNT1, NEXT)
```

In this example, if the value of KW is greater than or equal to 10, the next comparison, KW > 50, is made. However, for values less than 10, the PDF record is dropped and the next record is read. If the value of KW is greater than 50, the next PDF record is read. Only values between 10 and 50 inclusive are counted and stored in the field, COUNT1.

Action clauses may be written several ways, but have the same result. For example, the following Control File statements are functionally equivalent.

```
             KWH <= 100.0    T(COUNT1,NEXT) F(LINE2)
LINE2:       KWH <= 200.0    T(COUNT2,NEXT) F(COUNT3,NEXT)
             KWH >= 100.0    F(COUNT1)
             KWH >= 200.0    T(COUNT3)  F(COUNT2)
```

As illustrated in this example, use of the defaults can greatly reduce the size of your Control File. In general, set up your test statements so that you continue testing when they are true and go to the next PDF record when they are false.

DIM Statements

The DIM Statement is used to define the dimensions of a design and to specify the usage variable for which statistics are to be calculated. It must be used in conjunction with the CELL or STATS test statement options. Specific Control files that require a DIM Statement include: the Population Analysis Control File (TGB22A) for Multidimensional Population Analysis (PROC STARB200 or EXEC B200 — **Chapter 6**: **Assigning the Population to Cells and Calculating Population Statistics**); the Reporting Control File (TGB22A) for Single or Multidimensional Sample Selection (PROC STARB410 or STARB420; EXEC B410 or B420 — **Chapter 9**: **Selecting the Sample for a Single Dimensional Design** and **Chapter 10**: **Selecting the Sample for a Multidimensional Design**); and the Stratification Control File (TGB22A) for Single Dimensional Sample Selection (PROC STARB410 or EXEC B200 — **Chapter 9**: **Selecting the Sample for a Single Dimensional Design**). See Figure 6-2, Figure 6-4, Figure 9-2, Figure 9-3, Figure 9-4, Figure 9-5, Figure 10-4, and Figure 10-5 for examples.

Use this format to construct a DIM Statement:

DIMn a_1 a_2...a_n b value

where:

n	is the number of dimensions in the design. The value must be an integer between 1 and 7.
a_1 - a_n	is the number of strata in each dimension. There must be n values and each an integer between 2 and 9.
b	is the number of special cells in your design. You can specify up to 2 special cells. A value of zero **must** be used if there are no special cells.

value — is the usage variable in the Population Data File for which the statistics (population count, mean, and standard deviation for each cell) are to be computed in the program run. Be sure to spell the identifier for the usage variable the same way it was spelled in your Record Definition File.

Do **not** place a blank between DIM and n. Do place one or more blanks between all other parameters in the statement.

Format Statements — Tools For Creating Reports

Format statements enable you to specify the content and layout of a report. You can direct the program to print selected PDF files, add titles or notes, and insert blank lines or spaces to make the output more readable. Other formatting features includes tabulation, paging, and underscores for column heads.

Remember, the program executes a format statement when directed by the "PRINT n" option in an action clause. Consider the following example:

```
RATE = 'A1','A2','A3'  T(PRINT 1)
FORMATS:
1: 'CUSTOMER IS' CUSTID 'AND RATE IS' RATE SKIP(1)
```

For each customer in the PDF whose value for the variable RATE is A1, A2, or A3, the program will output a line in the report following the specified format. For example, for customer B003 who is in Rate Code A1, the program will output the following line: "CUSTOMER IS B003 AND RATE IS A1".

When creating your own format statements, keep in mind the following points:

- the same format statement may be referenced a number of times in the test statements.
- format statements must appear as a block after the test statements.
- they must be separated from the test statements by the keyword "FORMATS:".
- unlike test statements, format statements can be continued over multiple lines.

The total set of options available when creating a format statement is illustrated below.

```
n:   [variable [w.d]]  [literal]  [BLANK(s)]  [PAGE]   [SKIP(n)]
```

Each of these elements is described below.

- **n:** — identifies the format statement for reference by at least one PRINT option in the test statements. It can be any integer from 1 to 99 inclusive. The integer "n" must be followed by a colon (:) with no intervening blanks. Format labels need not appear in consecutive order.
- **variable** — identifies a field in the PDF record (as defined in the Record Definition File), the contents of which are to be written to the output file or report. You can have any number of variables in a format statement. There are two variables in the example format statement above: CUSTID and RATE.

  Unless otherwise specified (see "w.d" below), character data is printed in the report using the byte length specified in the Record Definition Program.

  The program automatically converts the contents of the field, no matter what its data type, to the EBCDIC equivalent before output.
- **w.d** — an optional field-specification of the form w.d, where "d" is the number of fractional digits and "w" is the total field width (integer digits plus decimal point plus fraction digits). The "w.d" field specifications may be used for character data as well, but only the field width is used. For instance, if you wished to print out just the first four characters in the variable field named "CUSTID", you would specify "CUSTID 4".

- **literal** — an optional character string, not exceeding 20 characters, used to input titles, field labels, or notes, or to insert blank spaces between fields in the report. You must enclose your entry in single quotes.

  If you wish to have more than 20 characters in a title, etc., you must break the characters up into subgroups. They will print continuously on the report. Example:

  'THIS IS A TYPICAL SU' 'BSTRING'

The following options are useful formatting aids:

- **BLANK(s)** — inserts "s" blank spaces in the output line. You can specify from 1 to 132 spaces inclusive. You must enclose the number in parentheses.
- **PAGE** — forces a page eject.
- **SKIP(n)** — executes a line return and skips "n" lines.

The options are executed in the order they appear within the format statement. For example:

```
1:  PAGE    SKIP(2)    BLANK(5)    'CUSTOMER ID'
            SKIP(0)    BLANK(5)    '___________'
            SKIP(2)    BLANK(5)    CUSTID 11
```

The above set of format statements causes a page eject, two line feeds, a tab of 5 spaces, and 11-character literal output, a line return with no skip, a tab of 5 spaces, another 11-character literal output (underscores), two more line feeds, a tab of 5 spaces, and a final 11-character field output. See **Creating a Report — A Step-by-Step Example** on page A-16 for additional information about formatting aids.

End-of-Program Statement

The End-of-Program Statement is a special type of optional format statement, which is not tied to a PRINT instruction in the test statements. All other format statements are executed while the PDF is being processed and PRINT options are encountered. The End-of-Program Statement is only executed when a STOP action occurs or the end of the PDF is reached. It is normally used to print out summary information and counter values.

The End-of-Program Statement appears after the regular format statements, and is separated from the regular format statements by the keyword "END:". All of the regular format statement options are available except the format label. Like a regular format statement, the END Statement may be continued over more than one line.

Following is a portion of an End-of-Program statement that prints a frequency distribution for Population Analysis:

```
FORMATS:
END:
'0.0'COUNT1 SKIP(1)
'5.0'COUNT2
SKIP(1)'
'10.0'COUNT3SKIP(1)
```

> **Important Note:** The keyword "END:" must be used as a component of a format statement. It is not used to indicate the end of a file. This will cause the program to abort.

File Statements

File statements enable you to rewrite entire PDF records to a "Selection File." You can also update selected PDF fields before output.

As explained, the program executes a file statement when directed by the "WRITE n" option in an action clause. Consider the following example:

```
STRATA > 1  F(WRITE 1)
STRATA > 2  F(WRITE 2)
STRATA > 3  F(WRITE 3)

FORMATS:
FILE:
1:  RAN# = RANDOM(152433, 0.300)
2:  RAN# = RANDOM(768961, 1.000)
3:  RAN# = RANDOM(596825, 1.000)
```

Depending upon the customer's stratum assignment, the program will insert a random number into the customer's RAN# field and output the resulting PDF record to a Selection File.

When creating file statements, keep in mind the following points:

- the same file statement may be referenced by more than one WRITE option.
- file statements must appear as a block after the format statements (including the optional End-of-Program Statement).
- they must be separated from the format statements by the keyword "FILE:".
- they can be continued over more than one line.

The general format of a file statement:

```
n:    [replacement-clause]  [replacement-clause]…
```

- **n:** — identifies the file statement for reference by at least one WRITE option in the test statements. It can be any integer from 1 to 99 inclusive. The integer "n" must be followed by a colon (:) with no intervening blanks. File labels need not appear in consecutive order.
- **replacement-clause** — Most replacement clauses are of the form:

  field = value

  In the above expression, "field" refers to the name of a PDF field, and "value" is a user-supplied constant: e.g., "STRATA = 1". The statement replaces the contents of the PDF field with the specified value. All values are converted to the data type of the PDF field (CHARACTER, INTEGER, REAL, PICTURE, LOGICAL, or PACKED) before the assignment.

Random Number Option

A special form of the replacement clause is used to generate a random number. It is used specifically in the Stratification Control File for Sample Selection (PROC STARB4120 or STARB420; EXEC 410 or 420).

The format for this clause is

RAN# = **RANDOM** (*seed* [,*cutoff* | ,1.0])

- **RANDOM** — the RANDOM option in a replacement clause generates a floating point number at random from a uniform distribution defined over the interval (0.0, 1.0). The

random number is stored in the PDF field named RAN#, which should have the REAL data type.

- **seed** — a random number seed that the program uses to initialize the random number process. Specify the seed as a 6-digit, odd integer between 0 and 999999.
- **cutoff** — the purpose of the cutoff is to limit the number of PDF records that must be sorted and reported in subsequent steps of the Sample Selection Program. The cutoff default of 1.0 places no limit on the output of PDF records.

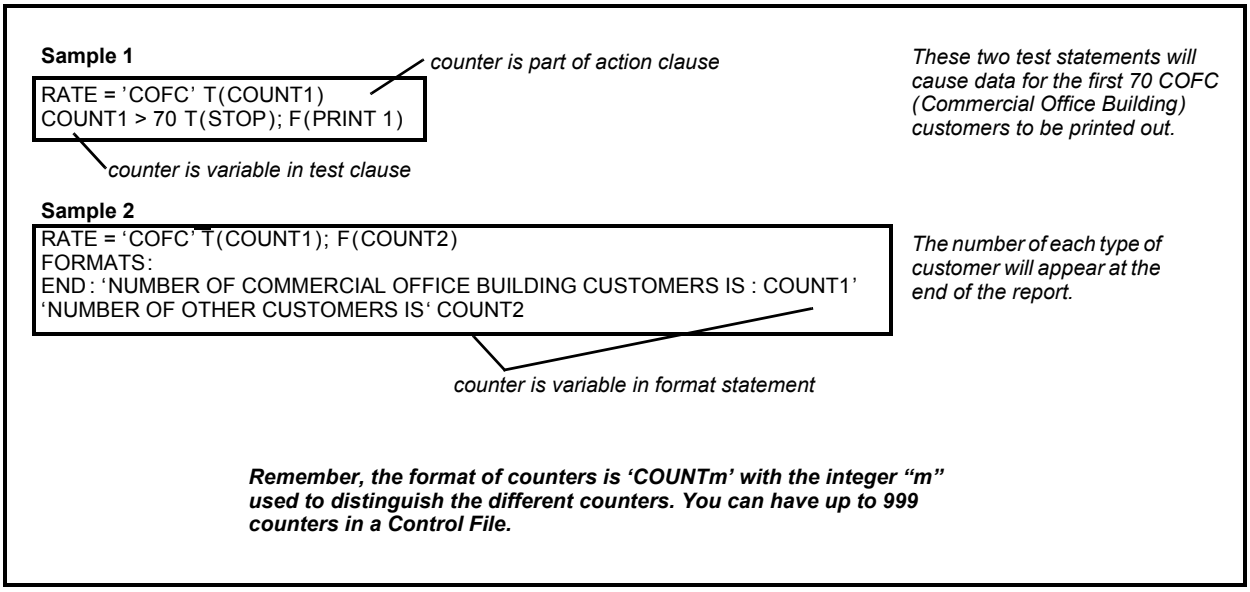
See "Step 1C — Create the Stratification Control File" in **Chapter 9**: **Selecting the Sample for a Single Dimensional Design** for additional information about the random number option.

Counter Variables

Counter variables enable you to keep a running count of the occurrence of specified conditions as the program applies your test statements to the PDF records. They can be used in several ways to accomplish different tasks:

- *as the specified action in true / false clauses* — to count the number of records matching or not matching your criteria.
- *as the variable in test clauses* — to initiate other actions when a certain count has been reached.
- *as a variable in format statements* — to output the number counted. Counter variables are treated as integer variables with a default field width of 8.

The following two samples illustrate the three different applications:



Creating a Report — A Step-by-Step Example

Probably the quickest and easiest way to learn the User Language is to actually apply it to a specific task. In this exercise, we will use elements of the language to create a custom report. Specifically, we'll apply the test statements to select customer records by desired criteria and the format statements to present the results in a custom-designed format. We'll pay particular attention to the formatting aids.

In our hypothetical example, we want to create a "Meter Installation Report" for each customer that was selected to participate in a load research study. Each report will show information about the customer that the Metering Department requires in order to install the load research meter, and it will have blanks for the installer to record information about the job. To get an idea of where we're headed, you can take a peek at the end result (Figure A-4).

Our imaginary sample design has three strata. Fifty customers are required for the first stratum, 65 for the second, and 60 for the third. The layout of our hypothetical Population Data File is described in the Record Definition File shown in Figure 3-1 (**Chapter 3**: **Creating the Population Data File and Record Definition File**).

The file we are going to create is the Reporting Control File for B410, Single-Dimension Sample Selection. The instructions in this example assume that the other input files required by the procedure have already been created; i.e., the Stratification Control File that assigns strata and random numbers to each eligible customer in the PDF and the Sort File that enables the programs to sort the customers by those numbers.

```
                /*  METER INSTALLATION REPORT  */
(1)             STRATA = 1   T(COUNT1)  F(STR2)                 /* SELECT FIRST 50 CUSTOMERS      */
(2)             COUNT1 <=  50 T(PRINT 1 , NEXT)                 /* IN STRATUM 1 & PRINT REPORTS   */

(3) STR2:       STRATA = 2   T(COUNT2)  F(COUNT3, STR3)         /* SELECT FIRST 65 CUSTOMERS      */
(4)             COUNT1 <=  65 T(PRINT 1 , NEXT)                 /* IN STRATUM 2 & PRINT REPORTS   */

(5) STR3:       COUNT3 <= 60  T(PRINT 1)                        /* SELECT FIRST 60 CUSTOMERS      */
                                                                /* IN STRATUM 3 & PRINT REPORTS   */

(6) FORMATS:                                                    /* BEGIN NEW PAGE FOR EACH        */
(7) 1:   PAGE                                                   /* CUSTOMER                       */

                                                                /* REPORT TITLE FORMAT            */
(8)
    BLANK(54)       'ABC ELECTRIC COMPANY'                                  SKIP(1)
(9) BLANK(47)       'RATE CLASS ' RATE ' (SMALL GENERAL SERV' 'ICE) '       SKIP(2)
    BLANK(51)        'METER INSTALLATION R' 'EPORT'                         SKIP(1)
    BLANK(51)       'WORK ORDER NO. :   ' ____ '                            SKIP(3)

(10)                                                            /* CUSTOMER INFORMATION FORMAT    */
    BLANK(2)        'CUSTOMER DATA :          '                             SKIP(1)
    BLANK(5)        'CUSTOMER ID :            ' CUSTID                      SKIP(1)
    BLANK(5)        'NAME :                   ' NAME                        SKIP(1)
    BLANK(5)        'STRATA :                 ' STRATA 6                    SKIP(3)
                                                                /* FIELDS FOR INSTALL INFO        */
(11)
    BLANK(2)        'INSTALLATION DATA :      '                             SKIP(2)
    BLANK(5)        'DATE INSTALLED :         ' '___ / ___ / 9__'           SKIP(2)
    BLANK(5)        'RECORDER SERIAL NO:      ' '_____',                    SKIP(2)
    BLANK(5)        'RECORDER LOCATION :      ' '_____'                     SKIP(2)
    BLANK(5)        'INTERVALS/HOUR :         ' '_____'                     SKIP(2)
    BLANK(5)        'METER LOCKING RING:      ' 'YES   /   NO   '           SKIP(4)

    BLANK(25)       '          CHANNEL 1      '
    BLANK(2)        '          CHANNEL 2      '
    BLANK(2)        '          CHANNEL 3      '                             SKIP(0)
    BLANK(25)       '_____ ',
    BLANK(2)        '_____ ',
    BLANK(2)        '_____ '                                                SKIP(2)
    BLANK(25)       ' METER NUMBER :____      ' '_____',
    BLANK(2)        '_____    ' '_____',
    BLANK(2)        '_____    '                                             SKIP(2)
    BLANK(25)       ' METER MULTIPLIER :      '
    BLANK(2)        '_____    ' '_____',
    BLANK(2)        '_____    '                                             SKIP(2)
    BLANK(25)       ' METER OFFSET : '
    BLANK(2)        '_____    ' '_____',
    BLANK(2)        '_____    '                                             SKIP(2)
    BLANK(25)       ' NO. OF DIALS :'
    BLANK(2)        '_____    ' '_____',
    BLANK(2)        '_____    '                                             SKIP(2)
    BLANK(25)       ' PULSE MULTIPLIER :'
    BLANK(2)        '_____    ' '_____',
    BLANK(2)        '_____    '                                             SKIP(2)
    BLANK(25)       ' PULSE OFFSET :'
    BLANK(2)        '_____    ' '_____',
    BLANK(2)        '_____    '                                             SKIP(2)
```

Figure A-4 Control File with Comments

Description

Note: Blank lines can be inserted between the statements as desired. They do not affect processing.

1. Our first task is to select the required number of customers in each stratum using test statements.

2. If the customer' value in the STRATA field is 1, the program will increment the variable COUNT1 by 1. If not, it will continue testing the customer with the test statement labeled 'STR2'.

3. For the first fifty customers counted in stratum 1, the program will execute the format statement labelled "1:" (PRINT statements initiate format statements with the corresponding label), and will restart testing with the next record in the PDF. Once 50 customers have been counted for Stratum 1, the program will ignore any other customers in that stratum; that is, it will execute the default for false action clauses F(NEXT).
4. If the customer's value in the STRATA field is 2, the program will increment the variable COUNT2 by 1. If not, the customer's value must be 3, so the program will increment COUNT3 by 1 and will go to the test statement labelled "STR3:".
5. Similar to the second test statement, except the program will execute the format statement labelled "1:" for the first 65 customers in Stratum 2.
6. Similar to the second and fourth test statements, except the program will execute the format statement labelled "1:" for the first 60 in Stratum 3. The NEXT option is not required in the true action clause as it was in the test statements above, because NEXT is the default option for the last test statement in the Control File.

Our second task is to specify how the resulting information is to be organized in the report using format statements.

7. The keyword "FORMATS:" is required between test statements and format statements. All of the following instructions in the file are a part of a block identified by the label "1:" (Format statements, unlike test statements, can run over more than one line). These instructions are initiated each time the program encounters a "PRINT1" option in the test statements while it is processing the PDF.
8. The first option, PAGE, begins a new page.
9. This line of instructions prints the first line of the page title. First, the printer will tab over 54 spaces, then print the "literal", that is, all characters including blanks between the apostrophes. It will then do a line feed.
10. The printer will tab over 47 spaces, then print the literal 'RATE CLASS' and the contents of the PDF field identified by the variable name RATE (as defined in the Record Definition File). Note the difference: characters and blanks enclosed in apostrophes will print in the report as you type them in the Control File; the variable name *not enclosed in apostrophes* will print the contents of the PDF field. Note also that the remainder of the title, "(SMALL GENERAL SERVICE)", had to be broken up into 2 subgroups. That is because literals can be no more than 20 characters long in the Control File. The title will print unbroken on the report.
11. This set of instructions will print field titles and customer information from the PDF. Note that the field specification option has been used to allot 6 spaces for printing the contents of the STRATA field. Without the field specification, the program prints out the number of spaces allotted to the field in the PDF.

Appendix B

Sampling Equations

This appendix documents the single-dimensional and multidimensional sampling calculations used in the Oracle Utilities Load Analysis Multidimensional Sampling Package. This material assumes a basic understanding of statistics.

Single-Dimensional Sampling Equations

Let:

L	=	number of strata in the sample design
h	=	index of the stratum
		Note: The range of h is from 1 to L for any given sample design
N	=	total population size
n	=	total sample size
N_h	=	population size of stratum h
n_h	=	sample size of stratum h
X	=	customer usage variable
x	=	sample mean of usage variable
x_h	=	mean of usage variable for stratum h
σxh	=	standard deviation of stratum h usage around the stratum mean
W_h	=	weight of stratum h
fpc_h	=	finite population correction for stratum h
Xst	=	sample mean across all strata
$\sigma_{x_{st}}$	=	standard deviation of $\bar{x}_{st}$
V	=	coefficient of variation (%)

- **Dalenius Hodges Method for Strata Breakpoint Determination**

f_j	=	population in segment j

i = index of the stratum

U value $U_j = j_i - j_{i-1}$ (segment length)

Cumulative value $cum\sqrt{uf} = \Sigma\sqrt{u_j}f_j$

Stratum Interval Length $K = \frac{cum\sqrt{uf}}{L}$

Stratum Breakpoints are at points j that are closest to interval lengths of K.

- **Determination of Total Sample Size**

Total sample size

$$n = \left| \frac{\left(\sum_{h=i}^{L} (w_h \sigma_{xh}) \right)^2}{\sigma_{x_{st}^2} + \frac{1}{N}\left(\sum_{h=i}^{L} W_h \sigma_{xh}^2 \right)} \right|$$

Population Stratum h Weight $W_h = \frac{N_h}{N}$

Finite population correction for stratum h $fpc_h = 1 - \frac{N_h}{N}$

Standard deviation of $\bar{x}_{st}$ $cum\sqrt{uf} = \sum_{j=1}^{i} \sqrt{u_j}f_j$

Coefficient of variation (%) $V = \frac{\sigma_{x_{st}}}{\bar{x}} * 100$

Variance of sample mean across all strata

$$\sigma_{x^{2}_{st}} = \sum_{\vartheta=1}^{L} \left(\left(\frac{w_h^2 \sigma^2_{xh}}{N_h} \right) fpc_h \right)$$

or

$$\sigma_{x^{2}_{st}} = (.01 * V * \bar{x})^2$$

- **Neyman Allocation for Strata Sizes Determination**

Sample Size for Stratum h

$$N_h = N \left| \frac{w_h \sigma_{xh}}{\sum_{i=1}^{L} w_i \sigma_{x_i}} \right|$$

Weight of stratum h

$$N_h = \frac{N_h}{N}$$

Standard deviation of stratum h's usage variable

$$\sigma_{x_h} = \left(\frac{\sum_{i=1}^{N_h} (x_{h_i} - \bar{x}_h)^2}{N_h - 1} \right)^{1/2}$$

Multidimensional Sampling Equations

Let:

L	=	number of cells in the sample design
C	=	index of the cell
		Note: The range of C is from 1 to L for any given sample design
N	=	total population size
n_m	=	total sample size based on usage variable m
N_c	=	population size of cell c
n_c	=	sample size of cell c
x_m	=	customer usage variable m
		Note: up to 7 usage variables are allowed
X_m	=	sample mean of usage variable m
$\overline{X}_{m,c}$	=	mean of usage variable m for cell c
$\sigma^2_{x_{m,cell}}$	=	standard deviation of cell c in usage variable
W_c	=	weight of cell c
fpc_c	=	finite population correction for cell c
$\overline{X}_{m,cell}$	=	sample mean of usage variable m across all cells
$\sigma_{x_{m,cell}}$	=	standard deviation of $\bar{x}_{m,\,cell}$

- **Determination of Total Sample Size**

Total sample size for dimension of usage variable m

$$n_m = \left[\frac{\left(\sum_{c=1}^{L} (w_c \sigma_{x_{m,cell}}) \right)^2}{\sigma^2_{\bar{x}_{m,cell}} + \frac{1}{N}\left(\sum_{c=1}^{L} W_c \hat{\sigma}_{x_{m,\,c}} \right)} \right]$$

Population cell weight

$$W_c = \frac{N_c}{N}$$

Finite population correction for cell c

$$fpc_c = 1 - \frac{N_c}{N}$$

Standard deviation of $\bar{X}_{m,c}$

$$\sigma_{x_{m,c}} = \sqrt{(\sigma_{\bar{x}_{m,c}})^2}$$

Coefficient of variation (%)

$$V = \frac{\sigma_{x_{m,c}}}{\bar{x}_m} * 100$$

Variance of sample mean across all cells

$$\sigma^2_{x_{m,c}} = \sum_{c=i}^{L} \left(\left(\frac{w_c{}^2 \sigma^2_{x_{m,c}}}{N_c} \right) fpc_c \right)$$

or

$$\sigma^2_{x_{m,c}} = (.01 * V * \bar{x}_m)^2$$

Note: A sample size is computed for each usage variable.

- **Neyman Allocation for Strata Size Determination**

Sample Size for cell c

$$n_c = n_m = \left(\frac{w_c \sigma_{x_{m,c}}}{\sum_{c=1}^{L} W_i \sigma^2_{x_{x,m,i}}} \right)$$

Weight of cell c

$$W_c = \frac{N_c}{N}$$

Sample Size for Stratum h

$$N_h = N \left(\frac{w_h \sigma_{xh}}{\sum_{i=1}^{L} w_i \sigma_{x_i}} \right)$$

Weight of stratum h

$$W_h = \frac{N_n}{N}$$

Standard deviation of cell c and usage variable m

$$\sigma_{x_{mc}} = \left(\frac{\sum_{i=1}^{N_c} (x_{c_i} - \bar{x}_c)^2}{N_c - 1} \right)^{1/2}$$

Note: Calculations use largest number of n_m for all m usage variables.

Index