

Defining Test Data Using Population Analysis

Clarence Wm. Jackson, CQA

Abstract:

Defining test data that provides complete test case coverage requires the tester to accumulate data that will satisfy all of the test cases in the requirements. Having access to production data eases the accumulation somewhat, but then any smart tester knows that testing should not be done in the production data. The next best situation is to completely copy production data, but this can be expensive depending on the size of the files and the amount of testing being committed. Another option would be to build test data based on the requirements, which may or may not match the varied data found in production data. This paper proposes using population analysis of production data as a tool to defining test data coverage. The tools available to provide statistical analysis, such as frequency of data element values and percentages to the total population, will provide measurable verification of the test data to the production data, ensuring better results of testing scenarios.

Introduction:

Test data is very important in ensuring that software performs as required. Test data is usually created by the testers using software specifications, which usually means that the software will meet the specifications. This approach may not test the total requirements of the software or how the software will be used in production. If the tester has no knowledge of the data and how it's used, vital requirements can be missed in testing. Since the test data is an important link in the software testing cycle, ways to ensure that the test data is adequate should be explored. Population Analysis of production data will reduce the risk of missing critical data types, and increase the reliability of the software when moved to production.

What is Population Analysis?

Population Analysis is simply creating reports that describe the type, frequency, and characteristics of data used by the application being tested. These reports will help in locating data and in defining the ranges and limits of data elements, lengths of fields, and the number of data values within each data element. The benefits of using Population Analysis are many, but the major benefit is to the tester in the following ways:

- Identification of codes and values being used in production which were not indicated in the specification or requirement document
- Unusual data conditions, like special codes, trigger values
- Provides a model for use in creating test transactions and test cases
- Provides a model for the type and frequency of transactions that should be created
- Helps identify incorrect transactions for testing error processing
- Documentation for use later in the project life cycle

There are three types of population analysis that can be performed to gather identifying information for testing. All the information gathered from the analysis should be documented. The three types of population analysis are:

1. File/data stores - to identify files and other input/output used by the software being tested
2. Screen - to identify screen displays that will be used or generated by the software
3. Field/data element - to identify characteristics and frequencies of fields/data elements

Usually the file/data stores are known, and the screens are straight forward. Of the three types, the Field/data element analysis is the most complex and time-consuming part of population analysis. That's where SAS comes in. Population analysis is best performed using software, and SAS can deal with any type of data encountered in any shop. SAS is designed for population analysis, and can compute all the statistics needed. There are so many PROCedures that will give you the information that it is a natural tool for population analysis. Other software that will provide analysis information are Audit Software, DataBase utilities, and CASE tools. But for field/data element analysis, SAS is the choice.

The steps for population analysis as presented in this paper are based on the Deming/Shewhart PDCA Quality Circle for continuous process improvement, and the Quality Assurance Institute's (QAI) process "WORKBENCH" model. The steps are:

1. Define the inputs for analysis (Plan)
2. Implement procedures (Do)
3. Check procedures (Check)
4. Produce deliverables (Act)

Before beginning, define what information (deliverable) is required from the analysis, create forms or other documentation templates. The documentation requirements for the three types of population analysis should be established in the planning stage. The documentation will assist you in evaluating the completeness of your analysis, and allow you to measure your progress.

Define The Inputs For Analysis

Production population analysis assumes that there is production data in some form, and that it represents the types of records that will be processed by the application being tested. Even if the application is brand spanking new, there is some type of data that can be used to load the test files. The types of files/data stores that meet the criteria for the software being tested are:

- Existing production files being used as is
- Existing production files for which there will be minor changes to the file
- Production files that contain the same or similar fields/data elements that will be included
- Existing manual files, such as paper invoices, forms, etc
- Files from other systems that contain the same or similar data elements

Implement Procedures

The major task is to identify the files that will be used for your analysis. Files and data stores must be located and defined. Flowcharts, data flow diagrams, system charts, and other

documents can be used to identify files within the scope of the testing project. To perform the analysis for files and data stores, you should execute the following three steps:

1. Identify file/data stores location, medium, organization of data, and other details
2. Identify the record types on the file/data stores, number of records, and sizes
3. Document the file/data stores characteristics

Screens display the data elements from the files/data stores in a human readable format. The application may need to reformat data before passing the data to the screen for display, such as in packed data. To perform the analysis for screens, there are two main steps that must be done:

1. Identify screens used, the data elements presented, and other details
2. Document screen characteristics

Field/data element analysis is the most complex of the analysis of population data. The analysis of the data is the most important, and time spent in this analysis will have a direct payback in improving the test data. The objective of this type of analysis is to describe the type, frequency, and characteristics of the data elements. It is this analysis that will be used to produce the test data. There are four main steps for field/data element population analysis. The steps are:

1. Identify each data element/field in the file/data stores
2. Document the characteristics of each data element
3. Analyze the population of data recorded in this data element area
4. Identify and document abnormal conditions

The analysis of the data should include the frequency, maximum, minimum, percentage of the population, and total values for each data element in each file/data store. This information should be documented.

Check Procedures

You should have mostly everything needed at this point, but it's a good time to review. Check your progress using your documentation template. Are you getting the information needed to perform a good test, and create good test data? Review abnormal conditions with those involved with the data for verification. All abnormal conditions should be tracked and resolved.

Produce Deliverables

The deliverables are the documentation which describe the population of the files, screens, and data elements. With the correct analysis, a question concerning the percentage of records having certain values is known and documented. The number of values within a given data element is known and documented. You now know what data values are required, and a much better set of tests can be performed. With this information, the test project should proceed with more assurance that the correct items will be tested. The documentation can be used for any of the following purposes:

- Supplement the requirements and specification documentation

- Development of the test plan
- Creation of test cases and scripts
- Creation of the test data
- Stress testing

How Can SAS Be Utilized To Perform Population Analysis?

SAS can be used in many ways to provide the for all types of population analysis reports though the many tools that is a part of the SAS System. SAS can be used for every type, or just for the more labor intensive analysis. In some cases, part of the file/data stores, screen, and data element analysis may be known, and provided to the tester. However, the analysis of the data element values should still be performed, and using SAS to do it will save many hours of manual checklisting. And SAS can be used to store the results of each type of analysis and report it.

For data element analysis, SAS will be able to read the data, and using your favorite PROC, such as FREQ, MEANS, TABULATE, etc, you can produce a listing of the frequencies of any and all data elements that are of concern. SAS can also load the test data stores using the values from the analysis, with a little coding using the DATA step programming statements.

Screen analysis can also be done using SAS, especially if it's in some format that can be read directly. Screens are such that tools may only be required to store the data from analysis.

The files and data stores can be analyzed manually, or using SAS. On the mainframe, PROC SOURCE, PDS, and other methods of getting file information from the operating system will give you plenty of information about the files. And if you are working with SAS data, PROC DATASETS, CONTENT, and other SAS PROCedures will provide everything from file information to data element metadata. For instance, PROC CONTENT will provide you with most of the identification information from SAS data files regarding each data element.

Conclusion

The use of population analysis for testing is a valuable tool in ensuring that software will function properly when moved to production. The best source of data for testing is production data. Why? It's data that is being used in the live process. To analyze the production data for the purpose of building test data is a sound method to verify software. Population analysis is a method to avoid the risk of missing obvious transaction types during the testing phase. The benefits of using the method are many.

SAS offers many tools that can be used to perform population analysis of production data. SAS can readily compute mean, median, mode, standard deviation, and other statistics by using a few BASE SAS PROCedures for field/data element population analysis.

Thank you for your time.

References

Jackson, Clarence Wm., *Using SAS To Help Manage Data*
Proceedings of the Sixth Annual South Central Regional SAS Users Group Conference, 1996

Perry, William E., *A Structured Approach to Software Testing*, Q.E.D Information Sciences, Inc.,
Wellesley, MA, 1983

Quality Assurance Inst., *QA/QC Solutions – Population Analysis: Creating Test Data From
Production Data*, The Process Warehouse, Quality Assurance Inst (QAI), Orlando, FL, 1997

Beizer, Boris, *Software Testing and Quality Assurance*, Van Nostrand Reinhold, NY, NY, 1984

SAS, SAS/AF, and the SAS System are registered trademarks of SAS Inst, Inc, in the USA and
other countries.

QAI and the Process Warehouse are registered trademarks of the Quality Assurance Inst, in the
USA and in other countries.

The author can be contacted via email as follows:

Clarence Wm. Jackson, CQA
CJac@compuserve.com
<http://ourworld.compserve.com/homepages/CJac>
cljacks@ci.dallas.tx.us (City of Dallas)