IJCSC International Journal of Computer Science and Communication Vol. 3, No. 1, January-June 2012, pp. 13-16

A STUDY ON SAMPLING TECHNIQUES FOR DATA TESTING

Manjunath T N[1], Ravindra S Hegadi[2] and Archana R A[3]

[1]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India, *E-mail: manjunath.tnarayanappa@rediffmail.com.*

[2]Department of Computer Science, Karnatak University, Dharwad, Karnataka, India, *E-mail: ravindrahegadi@rediffmail.com.*

[3]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India, *E-mail: archana.ra@rediffmail.com.*

ABSTRACT

Data quality has become increasingly important to many organizations to make any decisions on their business, as technology is evolved, the business demands to migrate the data from one platform to another due to various factors like portability, scalability and Security reasons, this paper emphasis on proposing model to do quality checks for huge database migrations using sampling techniques, the main idea of statistical inference is to take a random sample from a population and then to use the information from the sample to make inferences about particular population characteristics such as the mean, the standard deviation .Sampling saves money, time and effort. A sampling distribution is used to describe the distribution of outcomes that would observe from replication of a particular sampling method. The estimates computed from one sample will be different from another sample method. Different statistics have different sampling distributions with distribution shapes depending on (i) The specific statistic (ii) The sample size and (iii) The parent distribution. By understanding the relationship between sample size and the distribution of sample estimates. The variability in a sampling distribution can be used to reduce, by increasing the sample size. We expect this paper will help quality Analysts to analyze data quality of data migrations and explores effort and cost reductions during this process.

Keywords: Sample size, data testing, analysis and data migration.

1. INTRODUCTION

As technology is advanced, old systems has to be replaced with the new systems for portability and scalability, so data migration is common in each and every domain, in order to test the data migration process, which is tedious process and require lot of effort, because of huge volume of data. We can handle this using sampling techniques and new system can be proposed for conducting data migration testing. Data migration testing is needed when you move data from the old databases or one platform to a new database or another platform [14]. The old database is called the legacy system or the source database and the new database is called the target database or the destination system. A sample is a group of units selected from a larger dataset of population. By studying the sample, it is hoped to draw valid conclusions about the larger dataset. A sample is generally selected for study because the target dataset is too large to study in its entirety. The sample should be representative of the general target dataset. This is often best achieved by random sampling. Also, before collecting the sample, it is important that the author carefully and completely defines the population. Sampling is the process of selecting a small number of elements from a larger defined target dataset such that the information gathered from the small dataset will allow judgments to be made about the larger datasets. We explore the role of sampling in the data testing process, distinguish between probability and non probability sampling methods, understand the factors to consider when determining sample size and understand the steps in developing a sampling plan [6]. As a result by adopting techniques of Statistical Sampling, Statistical Quality Control (SQC) and Statistical Process Control (SPC), author should be able to define an approach to solve huge database migration quality assurance problems.

2. RELATED WORK

There are some efforts in the area of sampling methods for data testing, determining the proper sampling method is tedious task for those applying for data testing. Various literature is available on sampling methods are indirect or abstract regarding how, exactly, an appropriate sample can be constructed and what the sample can claim to reflect by Lindgren in 1993 and Rao in 2000. It is for this reason that this portion of the literature review de-mystifies the sampling process for those in the testing consortium. Through examining an appropriated sampling system constructed in consideration of the needs of state programs, this literature review offers a brief overview of the basic theories behind the sampling process, a step by step method states can use for creating a reliable sample, and a step by step method for determining whether a sample and its data is reliable. Ultimately, this section of the literature review should make compiling a reliable sample of a

population a relatively easy, straight forward process. A sample is the collection of people who were selected for a given research study by Frankfort-Nachmias and Leon-Guerrero, 1997. The most accurate and desired sample in any given situation is the probability sample by Sudman in 1976.Probability samples employ random sampling in order to ensure equal probability of selection method by Frankfort-Nachmias and Leon-Guerrero, in 1997 [6]. By compiling a complete list of a population, members of the population are systematically but randomly selected and then receive the survey or interview by Kish in 1965. After some of the research data has been collected, simple computations are applied to determine how many responses will be needed to ensure reliability by Baxter and Babbie in 2004 and Reinard in 2006. Once this number has been achieved, analysis determines whether a significant level of reliability has been achieved - and in the rare cases it has not, then more data is collected in an attempt to increase power in the study and allow for statistical significance by Kalton in 1998 [12].

3. OVERVIEW OF SAMPLING METHODS

It is incumbent on the researcher to clearly define the target population. There are no strict rules to follow, and the researcher must rely on logic and judgment. The population is defined in keeping with the objectives of the study. Sometimes, the entire population will be sufficiently small, and the researcher can include the entire population in the study. This type of research is called a census study because data is gathered on every member of the population. Usually, the population is too large for the researcher to attempt to survey all of its members. A small, but carefully chosen sample can be used to represent the population. The sample reflects the characteristics of the population from which it is drawn. Sampling methods are classified as either probability or nonprobability [6]. In probability samples, each member of the population has a known non-zero probability of being selected. Probability methods include random sampling, systematic sampling, and stratified sampling. In nonprobability sampling, members are selected from the population in some nonrandom manner. These include convenience sampling, judgment sampling, quota sampling, and snowball sampling. The advantage of probability sampling is that sampling error can be calculated. Sampling error is the degree to which a sample might differ from the population. When inferring to the population, results are reported plus or minus the sampling error. In nonprobability sampling, the degree to which the sample differs from the population remains unknown.

3.1. Random Sampling

Statistical sampling method in which every record within the target database has an equal likelihood of being selected with equal probability. Use random-number generator in an extract tool or data analysis tool or query tool. Determine the required DSS and number of records in the target database. Program the random-number generator to calculate a number between 1 and the total number of records in the target database. Select records where the number generated is less than or equal to the number of records required for the sample plus 1 or 2. For example, a database contains 923441 records. The required sample is 723 records. Program the random-number generator to calculate a number between 1 and 923441. Select records where the number of generated is less than or equal to 724 or 725. The greater number allows for the statistical chance that a smaller number of records will be selected than is actually required [6].

3.2. Systematic Sampling

Sampling in which every n^{th} record is selected. Select ratio based on the ratio of required DSS to total number of database records. Randomly select the first record to be chosen based on the ratio used. Systematic sampling is appropriate when the data population is ordered in a truly random sequence, and there is no bias in its ordering. Use this approach when random sampling of the desired database is not feasible.

3.3. Stratified Sampling

Sampling a database that has two or more distinct groupings, or strata, in which random samples are taken from each stratum to assure the strata are proportionately represented in the final sample. Use stratified sampling when the database being sampled is a distribution in records such that a small number of records exist for one subtype [6]. For example, a customer database has 100,000 customer records. If there are only 300 "preferred" customer, a sample size of 1000 could have the potential to include no records from this group. Stratified sampling seeks to assure the sample has an equivalent proportion of "preferred" customers selected, based on the stratification by the preferred customer code. The sample should contain approximately three preferred customers in a sample of 1000 records (300: 100000 = 3:1000).

3.4. Cluster Sampling

Sampling a database by taking samples from a smaller number of subgroups such as geographic areas of the population. The sub-samples from each cluster are combined to make up the final sample. For example, in sampling sales data for a chain of stores, one may choose to take a sub-sample of a representative subset of stores each cluster into a cluster sample, rather than randomly selecting sales data from every store. This technique may be used to represent all retail sales data only when the clusters truly represent the same relative kinds of data and the same relative process consistency. For example, if the files contain the same fields, the processes are the same, and the training and performance measure of information producers is consistent [5].

3.5. Two-Stage Sampling

Sampling from multiple files or database of the same data type, and then conducting a sample from the combined subset of group of data. This technique is used when collecting data randomly from distributed data files or databases or from different time periods, and then randomly selecting a final sample from the merged samples. This is appropriate when there are many different files or databases to sample and the size of the merged samples is larger than necessary for assessment. The first stage of sampling assures an adequate representation of data from each file/database or time period. Each sample of the first stage should be proportionate to its subpopulation (of records) so on group inordinately biases the final sample. The second stage of sampling assures adequate representation of the combined samples to represent the complete distribution of data [6].

4. MATHEMATICAL FORMULATION

Calculate an initial standard deviation of the target database records [9].

1. Create initial sample of records (ISR). It should be noted that statistical methods call for a minimum of 50 records to be examined, in order to draw any meaningful conclusions.
2. Compute Standard Deviation of ISR's using the following equation.

$$s = \sqrt{\sum d * d / (n-1)}$$

Where

s = the standard deviation of ISR.

d = the deviation of any record/row form the mean or average.

n = the number of records/rows in the sample.

$\sum$ = "the sum of"

3. Calculate Data Sample Size (DSS) using the equation:

$$n = ((z * s)/B)^2$$

Where:

n = DSS = number of records /rows to extract

z = a constant representing the confidence level desired. Indicates how much QA team is confident that the measurement of the sample is within some specified variation of the actual state of the data population. It is the degree of certainty, expressed as a percentage, of being sure about the estimate of the mean. There are statistical charts containing these constants. However, the most-used confidence levels and their constants include: 90%: $z = 1.645$, 95%: $z = 1.960$, 99%: $z = 2.575$.

s = the standard deviation of ISR. It is an estimate of the standard deviation of the data population being measured. This is the degree of variation of errors within the data population. The larger the variation, the larger the sample size required to get an accurate picture of the entire population. The smaller the variation, the smaller the sample size required.

B = the Bound or the precision of the measurement. The represents the variation from the sample mean within which the mean of the total data population is expected to fall given the sample size, confidence level, and standard deviation.

4. *Extract Data Sample Records:* Once we know the DSS, we can extract records or rows from the target database using the above listed sampling techniques and analyse the data and compare with source data and find the route cause analysis manually [19].

5. RESULT AND DISCUSSION

Some of the results observed in the below screenshots, shows that if we need more confidence level more samples are required this is selected based on confidence interval and population size. This sample size is used to extract the data from target database based on the sampling method used. We need to verify the following quality assurance metrics to ensure complete quality assurance of data migration process.

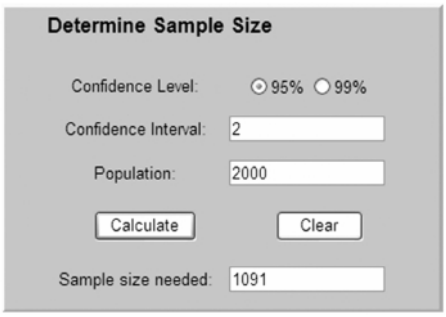

Figure 1: Data Sample Size with 95 Percent Confidence Level and 2 is the Confidence Interval for 2000 Population.

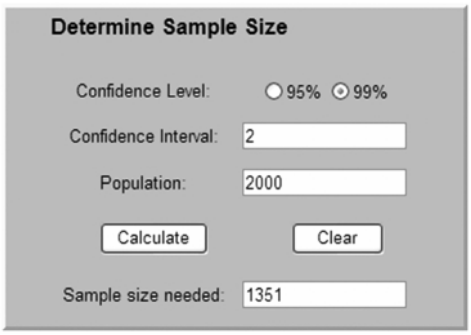

Figure 2: Data Sample Size with 99 Percent Confidence Level and 2 is the Confidence Interval for 2000 Population.

5.1. Quality Assurance Metrics used to Data Migration Testing

After extracting required number of records or rows from the target database by using suitable statistical sampling methods, then by computing the following data quality parameters for the desired samples and prepare defect matrix and Root cause [8].

Metric	*Description*
Accuracy-to the source	Is a measure of the degree to which data agrees with data contained in an original source. This measure is required to verify key source system data elements that did not change.
Validity-Business rule conformance	Is a measure of degree of conformance of data values to its domain and business rules. This includes: Domain values, Ranges, reasonability tests, Primary key uniqueness, Referential Integrity, This measure is required to ensure verification of key transformations.
Derivation Integrity	Is the correctness with which two or more pieces of data are combined to create new data. This measure is required to ensure verification of all computed fields which are present in target database and their calculation is correct.
Completeness-of values	Is the characteristic of having all required values for the data fields. This measure is required to ensure that a value is not missing from the target database.
Non-duplication-of occurrences	Is the degree to which there is a one-to-one correlation between records and the real world object or events being represented. This measure is required to ensure that duplicate representation of a single real-world object or event are not present in the target database.
Timeliness	Is the relative availability of data to support a given process within the timetable required to perform the process. This parameter is applicable after implementation in production environment.
Accessibility	Is the characteristic of being able to access data on demand. This parameter is applicable after implementation in production environment.

6. CONCLUSION

Data Migration from different legacy systems with huge data can be easily validated with the proposed study using samples, using the proposed method can reduce cost of testing effort and time and it is easy very a simple method for any data migration validations. Competition requires timely and sophisticated analysis on an integrated view of the data. This Paper describes as easily implemented, non-disruptive, scalable and comprehensive foundation for capturing data quality in the data migration work.

REFERENCES

[1] Ralph Kimball, "The Data Warehouse ETL Toolkit", Wiley India Pvt Ltd., 2006.

[2] KV.K.K Prasad,"Data Warehouse Development Tools", Dreamtech Press, 2006.

[3] W. H. Inmon, "Building the Data Warehouse", Wiley; 3rd Edition March 15, 2002.

[4] Arshad Khan, "SAP and BW Data Warehousing", Khan Consulting and Publishing, LLC, (January 1, 2005).

[5] Potts, William J. E., "Data Mining Using SAS Enterprise Miner Software", Cary, North Carolina: SAS Institute Inc (1997).

[6] SAS Institute Inc., (1998), "SAS Institute White Paper, From Data to Business Advantage: Data Mining", *the SEMMA Methodology and the SAS® System*, Cary, NC: SAS Institute Inc.

[7] G. John and P. Langley, "Static Versus Dynamic Sampling for Data Mining", *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 367-370, AAAI Press, Menlo Park, CA, 1996.

[8] Microsoft® CRM Data Migration Framework White Paper by Parul Manek, Program Manager Published: April 2003.

[11] Jaiwei Han, Michelinne Kamber, "Data Mining: Concepts and Techniques".

[12] "Data Migration Best Practices NetApp Global Services", January 2006.

[13] "DATA Quality in Health Care Data Warehouse Environments Robert L. Leitheiser", University of Wisconsin-Whitepaper.

[14] Ravikumar G K, Manjunath T. N, Ravindra S. Hegadi, Umesh I.M, "Cross Industry Survey on Data Mining Applications", *IJCSIT*, **2(2)**, 2011, pp. 624-628.

[15] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K., "A Survey on Multimedia Data Mining and Its Relevance Today", *IJCSNS*, **10**, No. 11, pp. 165-170.

[16] John Hess, "Dealing With Missing Values in The Data Warehouse" *A Report of Stonebridge Technologies*, Inc-1998.

[17] Manjunath T.N, Ravindra S Hegadi, Ravikumar G K., "Analysis of Data Quality Aspects in Data Warehouse Systems", *(IJCSIT) International Journal of Computer Science and Information Technologies*, **2(1)**, 2010, pp. 477-485.

[18] Jaideep Srivastava, Ping-Yao Chen, "Warehouse Creation-A Potential Roadblock to Data Warehousing", *IEEE Transactions on Knowledge and Data Engineering* January/February 1999 (**11**, No. 1), pp. 118-126.

[19] Manjunath T N, Ravindra S Hegadi, Mohan H S, "Automated Data Validation for Data Migration Security", *IJCA Online*, **30**/number 6/3642-5088: ISBN: 978-93-80864-89-0.