A Tool Generating Dataset for Decision Making

Kuldeep Arora
Research Scholar
Punjab Technical University, Jalandhar
Punjab, India

Jaiteg Singh, PhD.
Associate Professor
Chitkara University, Punjab,
India

ABSTRACT
Researcher and policy makers perform many of the survey techniques for collecting data to identify their findings and decisions. The aim is to demonstrate a tool TDG that generates the synthetic dataset for researchers and policy makers where initials estimates are not available. In this paper we are discussing the approaches & key issues in generating the synthetic dataset using TDG ensuring researcher & policy maker need not to go to the actual data.

Keywords
Synthetic Dataset, Decision Making, Research, Survey Techniques.

1. INTRODUCTION
In these days' research scholars, faculty members, employees of an organization, policy makers and statistical agencies use many of the survey techniques like observation, interview, and questionnaire etc. for collecting data. They deal with the economic and socio-economic data. Economic data is the historical records of economic system of an organization, country etc. used to formulate new policy. On the other side socio-economic data represents the conditions, attitudes, thinking and quality of life of the people. The aim is to generate the synthetic dataset for decision making using TDG.

Use of synthetic dataset is proposed by the Donald Rubin in 1993. Using Original Dataset has the limitations of privacy and disclosure. The use of synthetic dataset minimizes the disclosure risk. The Act Data Protection enforces the statistical agencies and individuals to keep the responses of survey to themselves. Authors suggest using synthetic data for decision making and research. With the use of synthetic data in research, researcher saves lots of time & expenses.

The synthetic data sets can be used in many ways such as Full Synthesis and Partial Synthesis. Full synthesize means synthesize of all variable for all respondents and Partial Synthesize means synthesize of a subset of variables for a subset of records.

Generally there are five basic type of questions included in any of the survey technique i.e. multiple choice, categorical data, likert scale, numerical and ordinal. This paper begins by discussing the shortcoming of current data collection techniques. Then introduce the approaches and issues to generate synthetic data using TDG.

2. OBJECTIVE OF STUDY
The authors are working on development of a tool that generates the synthetic dataset for researchers and policy makers. The objective is to generate realistic synthetic dataset that must preserve anonymity to the original data. The ultimate objective is to generate synthetic dataset that must be used for the public release with satisfying the statistical disclosure limitation using TDG. It may also happen that initial estimates are not always available so the use of synthetic dataset is the only choice for the policy makers. Authors put their efforts to generate the synthetic dataset for policy makers by responding synthetically the five close-ended question categories. The randomizers are provided in the tool TDG such as Random Integer, Library, List, Weighted List etc. In this paper authors describe how the tool TDG used to generate the synthetic dataset to respond the basic type of questions such as multiple choice, likert scale, numerical, ordinal etc.

3. CURRENT TRENDS FOR DATA COLLECTION
Data Collection is a process that must be performed for research and decision making. Data Collection is an important part of the research and decision making. Conclusions are find out on the basis of collected data because collected data is primary source of research. Following figure presents the picture for current data collection techniques.

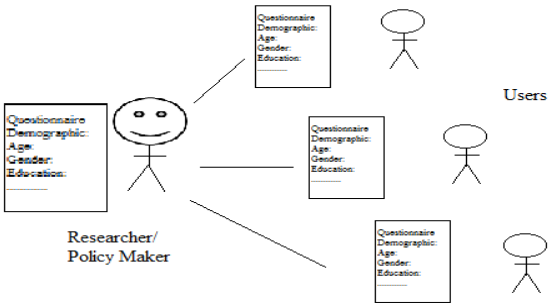


Figure-1: Current Scenario in Data Collection

There are two types of Data is collected Primary Data and Secondary Data. Primary Data is the data collected for the

first time and Secondary Data is the data that which already have been collected and availed by the statistical agencies.

During research and decision making researcher and policy makers has to perform the following data collection techniques:

3.1 Observation
In this method researcher goes to the field and collects information by observation. It is the method to collect data for gathering descriptive information.

3.2 Interview
Interview is an alternative method of collecting data. Interviewer asks questions orally & records the answer by respondents. It is a good method to gather information. Interview is the source of qualitative information. This method will produce better data on sensitive issues

3.3 Questionnaire
Questionnaire is a method for data collection. Questionnaire is an effective method for obtaining data from large audience. It is a method that provides proper authenticated data. Quantitative data will produced by distributing questionnaire with closed questions that offers respondents to choose answer, not expressing in writing. It is the best source of Qualitative and Quantitative data.

4. PROJECT IDEA
There are some of the shortcomings of these Survey techniques that lead to the use of synthetic data and idea for the research and decision making

1. These survey or data collection technique usually have low response rate.
2. These techniques are more expensive and time consuming.
3. It may happen that many of the respondents are unwilling to respond, that will suffer the results.
4. It is always not possible to locate the respondents with equal probability. i.e. inequality in response.
5. Careful framing and phrasing of the survey questions are necessary. Symantec difficulties are the major problem.
6. Survey methods require lot of expertise in collecting data. These techniques require expertise in properly design the survey, collecting the responses and processing the data.
7. The process of entering, editing and analyzing the collected data is the subject part of the research. It also requires careful handling of data. It is very time consuming process. Data entry is performed manually that may leads to error also.
8. The major shortcoming of the survey techniques is to maintain the privacy of the respondents.

5. CLOSE-ENDED QUSETIONS CATEGORY
Any of the Survey designed in two ways: Close-ended & Open ended. Generally Close-ended have five basic types of questions included in any of the survey technique. Authors discussed how the synthetic dataset can be generated using TDG for close ended survey. There are the generally five basic type of questions asked in any of the survey of close-ended questions.

5.1 Multiple Choices
For e.g. Which Flavor you like the most in ice cream?

- Vanilla
- Strawberry
- Black current
- Butter scotch

5.2 Categorical
For example, what is your marital status?

- Married
- Single

5.3 Likert Scale
For example, how much you interested in buying ready-to-eat food products?

(Strongly Disagree) 1 2 3 4 5 (Strongly Agree)

5.4 Ordinal
A ranking indicates the importance assigned by a participant. For example, please rank the following products from best (1) to worst (5) according to your choice

- Lays Chips
- Lehar Kurkure
- Stop Not
- Bingo
- Fun Flips

5.5 Numerical
When the number is answer?
For example, what is your age?

6. EXPERIMENTAL SETUP
TDG (Test Data Generator) is a tool that generates synthetic dataset for decision making developed in Microsoft Visual Studio compiler environment for all major relational data sources including MSSql Server, MySql, Oracle, Desktop Files such as delimited file, excel file & XML file etc. Authors demonstrates how the TDG generates the dataset for these 5 close-ended questions

The figure-2 showing our approach to generate synthetic data

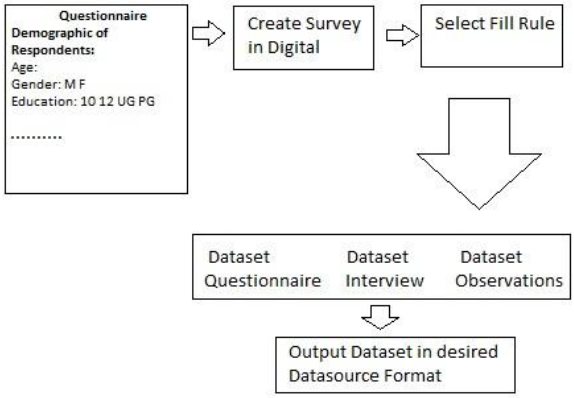


Figure-2: Our approach to generate data

The entire process is in 4 stages:

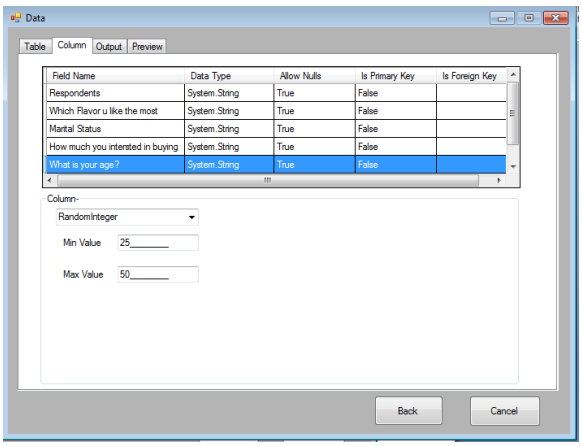
1. Design the survey
2. Set the fill rule (i.e. randomizer)
3. Set the limit of dataset to be generated
4. Output the dataset into desired format such as csv, excel etc.

Stage one is to design the survey in digital format using create schema feature of TDG. If the user has already survey in digital format he must connect to that source using Connect to a data source feature of TDG.

Once user is connected with the survey then in the second stage he must select the fill rule according to question type. The fill rules provided by TDG are Library, List, Weighted List, and Random Integer, Random Ordinal.

Random Integer is a randomizer that generates random numbers between limits. One challenge is to generate the responses for numerical type question category. Random integer randomizer is most suitable for this type question category.

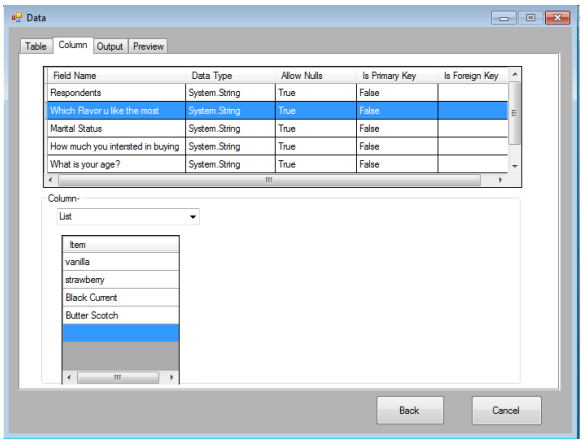
The following snapshots showing how the different randomizers used for different category type questions included in the survey techniques.



Snapshot-1: Showing Random Integer for a numerical type question

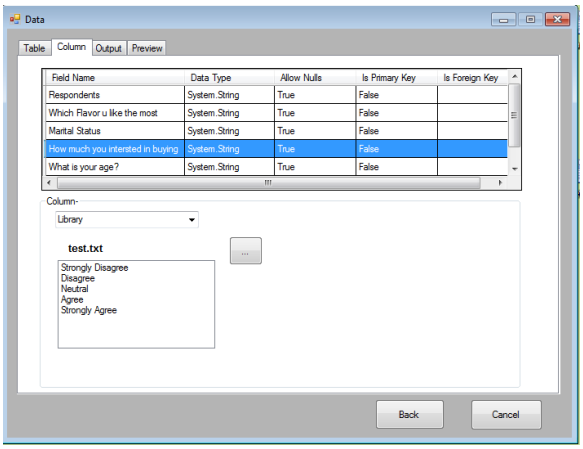
Another challenge is to generate synthetic dataset for question types – Multiple Choice, Likert Scale and Categorical. The Randomizer can be used for these type are List, Library, and Weighted List.

List Randomizer is the fill rule that select responses randomly from the list created by the user.



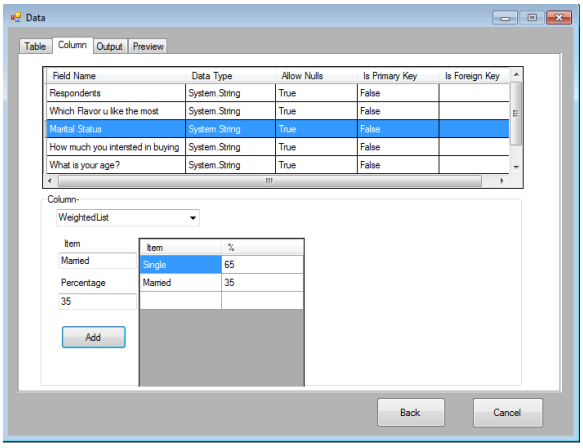
Snapshot-2: Showing List for a multiple choice type question

Library Randomizer is the fill rule that select responses randomly from user created library of text type.



Snapshot-3: Showing Library for a Likert-Scale type question

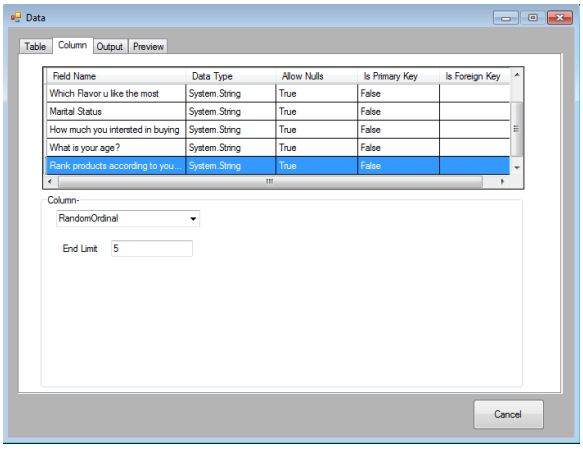
Weighted List is the Fill rule that generates responses according to weights set by the user. This randomizer is very useful for the statistical System.



Snapshot-4: Showing Weighted List for a Categorical type question

Weighted list is very useful randomizer to generate responses with specific weights according to percentage.

Another challenge is to generate the synthetic dataset for Ordinal type question. Random Ordinal Randomizer is specially created for this purpose. Random Ordinal randomizers generate responses in ordinance randomly.



Snapshot-2: Showing Random Ordinal for Numerical type question

In the third stage user set the limit of dataset to be generated for survey. Next and the last stage user can generate the dataset in desired data source format such as csv, text, Excel etc.

Following Snapshot showing result from Excel File generated dataset for all five type question category by TDG.

Respondents	Which Flavor u like the most	Marital Status	How much you intersted in buying	What is your age?	Rank products according to your choice
1	vanilla	Married	Agree	55	21345
2	vanilla	Single	Agree	54	42315
3	black current	Single	Strongly Disagree	29	12345
4	vanilla	Married	Agree	22	42315
5	butter scotch	Single	Neutral	31	12345
6	vanilla	Single	Agree	23	32145
7	butter scotch	Married	Neutral	25	32145
8	vanilla	Single	Disagree	41	52341
9	strawberry	Single	Neutral	49	32145
10	butter scotch	Single	Disagree	40	12345
11	vanilla	Married	Disagree	58	12345
12	vanilla	Single	Strongly Disagree	42	42315
13	black current	Single	Neutral	23	32145
14	butter scotch	Single	Agree	37	32145
15	butter scotch	Married	Agree	31	21345
16	vanilla	Single	Disagree	56	42315
17	black current	Married	Neutral	33	21345
18	vanilla	Single	Neutral	56	42315
19	black current	Single	Disagree	46	52341
20	black current	Married	Disagree	30	21345
21	strawberry	Single	Agree	20	42315
22	butter scotch	Married	Strongly Disagree	56	42315

7. ACKNOWLEDGMENTS

I would thanks to my guide Dr Jaiteg Singh who gives me pleasure to contribute in the research. I am always thankful to him for this.

8. REFERENCES

[1] Jaiteg Singh and Kawaljeet Singh, 2008,"Designing a customized Test Data Generator for effective Testing of a Large Database", In ICACTE

[2] Raghunathan, Reiter, Rubin, Vol 19, 2003, "Multiple Imputation for Statistical Disclosure Limitation", In journal of official statistics

[3] Daniela Ichim, 2010, "Quantile-based Bootstarp methods to generate continuous Synthetic data" In ACM

[4] Patrick Graham and Jim Young, 2008, "Methods for creating Synthetic data", Report.

[5] Little, R. J. A., *1993* "Statistical analysis of masked data",.*Journal of Official Statistics*

[6] J. Gray, P. Sundaresan, S. Englert, K. Baclawaski, and P.J. Weinberger, 1994, "Quickly generating billion record synthetic databases". In SIGMOD