**Homework 2. Selection of PPS Samples and Relevance of Sampling Statistics**

**MSDS 6370**

**Submitted by Hari Narayan Sanadhya**

**Objective:**

* For the student to learn how to select a probability proportional to size (PPS) sample and to form estimates of the total using a PPS sample.
* For the student to explore the relevance of sampling statistics in a time of Big Data.

**Introduction**

Asynchronous **Unit 2** included a discussion PPS sampling and how to form estimates with this sampling method. This assignment builds on that discussion and extends it by selecting a sample PPS. For a big picture view, the assignment includes reading and discussing two documents about the relevance of sampling in a world with Big Data.

**Data for PPS Sample**

In this assignment, you will use data from an audit of a health care provider. The objective of the audit was to estimate the total overpayment by an insurance company to a medical provider. The data does not include any personal information.

The file HW2\_data.xlsx contains the data that you will download. In the encounter sheet, each record in the file represents a patient’s encounter by the health care provider and contains patient number, the amount the provider was allowed to charge, the amount that was paid, and the difference. We are going to select a simple random sample with replacement (SRSWR) of encounters. As you will see most patients have more than 1 encounter so the sample we select can contain 2 or more encounters for a patient.

Selecting the sample in this manner creates a PPS sample that is not representative of the population as it is. A patient’s number of encounters affects the patient’s probability of selection and therefore is proportional to the patient’s size for the purpose of our sampling procedure. Usually too many patients with a large number of records (large units) will be selected, and too few patients with a small number of records (small units) are selected. We need to *downweight* units that occur too frequently (large units), and *upweight* those that occur too infrequently (small units). The appropriate weight of the unit equals the reciprocal of its size, which, in this case, is a patient’s number of encounters.

The sheet patient contains the patient number and number of encounters, which is the number of rows for the patient in the encounter sheet.

By the way, a cluster design with stratification was used in the real audit. We will discuss this type of design in detail later in the course.

**Exercise 1**

You may use the software of your choice in this exercise. The following describes the steps using Excel.

Select a PPS sample of encounters of size 15.

1. Assign each a random number to each patient in the encounter sheet, which is sorted by patient number. In Excel, you can label a column ‘Random Number’ and assign a random number to the record by using RANDBETWEEN(a,b), where a and b indicate the range for the random numbers. You can set a = 1, but since there are 2,471 rows of data (not including the headings) in the file, b =2,471. Setting b equal to a number higher than 2,471 also will work.
2. Sort all columns by the random number. (You may have to perform Copy-special-values-Paste on the random numbers before sorting.)
3. Take the first 15 records as your random sample.
4. Create an Excel file, called HW3\_sample.xlsx with just these 15 records from the encounter sheet. Name the sheet “sample”.

Merge the number of encounters onto each record in the sample.

1. Each record in the sheet patients has the patient number and the patient’s number of encounters. The patient sheet has 64 records.
2. For each record in HW2\_sample.xlsx, create a new column titled ‘number of encounters’ and record the number of encounters for each patient from the patient sheet.

Calculate the weight for each record in the sample

1. In the sample sheet, add a new column titled ‘weight.’ This column will contain the record’s weight that will be used in estimation.
2. The number of records for a patient is the patient’s measure of size for our PPS sampling procedure. The probability of selection of each patient on each draw is (# of encounters for that patient)/2471. (Note that 2471 is the total number of encounters in the file.)

Therefore, the weight for each patient would be the reciprocal of the selection

probability, or wi =2471/(# of encounters for that patient)

1. Record the data you have in the sample sheet in Table 1.

Table 1. Data for PPS sample of size 15

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sample Unit** | **Patient Number** | **Amount Paid** | **Amount Allowed** | **Difference** | **Size** | **Weight** |
| 1 | 10258 | $0.00 | $0.00 | $0.00 | 27 | 91.51852 |
| 2 | 3086 | $8.39 | $8.39 | $0.00 | 16 | 154.4375 |
| 3 | 3090 | $144.77 | $144.77 | $0.00 | 23 | 107.4348 |
| 4 | 232 | $0.00 | $0.00 | $0.00 | 51 | 48.45098 |
| 5 | 1479 | $85.00 | $85.00 | $0.00 | 38 | 65.02632 |
| 6 | 10265 | $0.00 | $0.00 | $0.00 | 40 | 61.775 |
| 7 | 3089 | $121.28 | $0.00 | $121.28 | 35 | 70.6 |
| 8 | 11422 | $144.35 | $0.00 | $144.35 | 238 | 10.38235 |
| 9 | 10595 | $177.00 | $118.00 | $59.00 | 59 | 41.88136 |
| 10 | 3142 | $143.32 | $143.32 | $0.00 | 8 | 308.875 |
| 11 | 10263 | $183.96 | $183.96 | $0.00 | 26 | 95.03846 |
| 12 | 10266 | $26.37 | $26.37 | $0.00 | 21 | 117.6667 |
| 13 | 4 | $130.20 | $0.00 | $130.20 | 317 | 7.794953 |
| 14 | 3147 | -$99.34 | -$99.34 | $0.00 | 75 | 32.94667 |
| 15 | 11422 | $188.48 | $189.52 | -$1.04 | 238 | 10.38235 |

1. What formula is appropriate for estimating the total Amount Paid using the PPS sample in Table 1? (See asynchronous session 2)

Estimated Mean =

where

wi = weight of the ith instance

yi = variable of interest of the ith instance

2. Estimate the total amount paid, the total amount allowed, and the total difference.

**Solution**: -



Estimated total amount paid

= (Sum of (amountPaid \* Weight)) / (Sum of Weights)

= 104403.32/1224.21091

= $ 85.28

Estimated total amount allowed

= (Sum of (amountAllowed \* Weight)) / (Sum of Weights)

= 90867.15/1224.21091

= 74.23

Estimated total difference

= (Sum of (difference \* Weight)) / (Sum of Weights)

= 13536.17/1224.21091

= 11.06

**Exercise 2**

The following 2 documents explore issues surrounding the relevance of sampling when Big Data is available:

* “Is Sampling Relevant in the Time of Big Data” debate on a discussion board on Stackexchange.com.

<http://stats.stackexchange.com/questions/35971/is-sampling-relevant-in-the-time-of-big-data>

* “The Hidden Biases in Big Data” by Kate Crawford, *Harvard Business Review*, 2013.

<https://hbr.org/2013/04/the-hidden-biases-in-big-data>

1. Write a brief summary of the main issues identified in “Is Sampling Relevant in the Time of Big Data.”

In the world today, where Data is being produced at astronomical rates, Big Data and its technologies have provided analysts capabilities to explore and analyze these datasets in very economical and efficient way. Given these capabilities, the discussion talks about if Sampling is still relevant or not. Though we have huge datasets with very large number of attributes, we have issues with the dataset still being representative of the population or not. If the dataset is representative of the population then we can ignore sampling and directly use the data for analysis and model building. If not, which almost always holds true, sampling from the dataset would be of utter importance as only based on the untouched samples of the data can the performance of the model can be evaluated. Methods like Cross Validation, k-means clustering, random forests use sampling behind the scenes without the user explicitly perform sampling.

With the increasing capability to store and analyze all the kinds of information, proper variable's selection for model building is getting harder to determine. The more the number of attributes we have in the dataset, its difficult to obtain which variables contribute more towards the variance in the variable of interest. This is because with substantial number of variables, the variance in the variable of interest is more likely to be closely distributed among the variables. Also, the deep knowledge of all the attributes in the dataset and its impact on the variable of interest is not likely to be known by everyone. So, the challenge here is micro-segmentation, i.e. pulling out the right details i.e. attributes that are relevant.

Apart from the issue of whether big data being representative of the actual population or not, the other issue is the validity of the data due to time dimension. With big data, if the statistical inference is aimed to support future predictions, then the population must be understood to extend into the future. Extrapolation of the results from the current population may be true only for some period as if the data has time dimension, the data is likely to show variance with time.

Also with big data, if proper sampling is not done but instead the model is build using the entire dataset, the non-sampling errors and biases are usually much bigger than the sampling error and biases especially while measuring opinions.

In a nutshell, sampling is very useful when the population to be studied is very large and you are interested in the macroscopic properties of the population. 100% checking (Big Data) is necessary for exploiting the microscopic properties of the system. This is true only when the big data is representative of the population not just in terms of existence of data from everyone part of the population but also the distribution of the data. Example if the task is to analyze the tweets to study the impact of a change in tax structure by the state For this, it might be possible that the tax changes made were more in favor of rural population then urban population who generally have high access to the internet, the analysis is likely to show that the tax change is not having a positive feedback from the population as tweets from people in urban areas would be far more then those from rural areas.

2. Write a brief summary of the main issues identified in “The Hidden Biases in Big Data.”

The article discusses about the validity of the idea embraced by Former Wired editor-in-chief Chris Anderson, “with enough data, the numbers speak for themselves.”.

With big data, as we have millions and millions of records in the dataset, this statement if true, would mean that we know all the possible information that the data can show. This is not true as the data collected is a result of human design, so the data is not objective. We are designing the process to collect the data and collecting what we need, and we can, and then interpreting based on the data collected. Hidden bias present at both the collection and analysis stage present considerable risks of these interpretations being true. The hidden bias includes the representativeness of the actual population in the data (example the data collected from the social media is not representative of the actual human population as not everyone is having access to internet and those who have, may not be using social media, and those using, may be not interested/commenting about the issue of interest that the analyst is analyzing.)

Secondly the geographical and the cultural variations also cause bias in the dataset. Places, like people, have their own individual character and grain making the dataset to be biased towards geographies/cultures. Signal issues and the reachability of the signal generating device to the desired population (like signals problems on smart phones and its usage by everyone is always desired if data from smart phone is to be analyzed) are other means by which bias can be introduced into the dataset.