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

**MSDS 6370**

**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 | 3147 | 0.00 | 0.00 | 0.00 | 75 | 32.9467 |
| 2 | 18 | 135.38 | 135.38 | 0.00 | 53 | 46.6226 |
| 3 | 3149 | 10.84 | 10.84 | 0.00 | 38 | 65.0263 |
| 4 | 3149 | 80.00 | 47.00 | 33.00 | 38 | 65.0263 |
| 5 | 4 | 10.84 | 0.00 | 10.84 | 317 | 7.7950 |
| 6 | 4 | 2.10 | 0.00 | 2.10 | 317 | 7.7950 |
| 7 | 11420 | 0.00 | 0.00 | 0.00 | 39 | 63.3590 |
| 8 | 10598 | 27.69 | 27.69 | 0.00 | 34 | 72.6765 |
| 9 | 10598 | 36.96 | 36.96 | 0.00 | 34 | 72.6765 |
| 10 | 10265 | 21.44 | 21.44 | 0.00 | 40 | 61.7750 |
| 11 | 3090 | 0.00 | 0.00 | 0.00 | 23 | 107.4348 |
| 12 | 3090 | 39.93 | 39.93 | 0.00 | 23 | 107.4348 |
| 13 | 230 | 20.62 | 21.44 | -0.82 | 59 | 41.8814 |
| 14 | 10600 | 0.00 | 0.00 | 0.00 | 23 | 107.4348 |
| 15 | 11423 | 214.99 | 214.99 | 0.00 | 96 | 25.7396 |

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

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

**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.”

The discussion pivots back and forth about the importance of sampling in the age of Big Data. As more and more information is collected and machines are powerful enough to handle larger swaths of information the argument is posed if sampling is relevant anymore. The short answer is it depends.

The key point is Big Data opens avenues that sampling will involve to. There are situations where it is more computationally expensive to run all the information through an algorithm with little improvement than if sampling methods were used. It is a case by case situation.

In addition, sampling has been an integral part in the development of new machine learning algorithms. Cross validation uses sampling methods to prevent models from overfitting to a dataset allowing for more generalized prediction models.

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

The article notes the dangers of limited scope in the Big Data era by collecting information from certain mediums. The notion that Big Data will allow numbers to speak for themselves is an overreach because it assumes that data collected is the correct data for the insights desired.

Using Twitter data to determine the impact of a hurricane is improper because it assumed that severity could be assess from those with an operating cell phone with service and available time to tweet about the hurricane. In reality, most affected by the storm were busy weathering the storm or without cell service to express their concerns.

**Exercise 2.1**

*A researcher wishes to study key business leaders to learn their opinions about issues facing the metropolitan area. Define this population in specific operational terms.*

The population is key business leaders within the metropolitan area. There is no distinction to what qualifies a “key business leader” from a “business leader” but boundaries can be arbitrarily determined. For instance, the researcher can frame the population to individuals who work for a business within the metropolitan area and has received an award from a civic organization for their community involvement.

**Exercise 2.2**

*The municipal government of a “college town” wishes to survey area residents regarding their park and recreation needs. Define this population in specific operational terms. Should children be eligible to respond? People who live outside the city boundaries? Students at the local university who have access to university facilities? Students who live in dormitories? Students in fraternities or sororities? People in the local jail? People in a homeless shelter? People in nursing homes?*

The population is individuals living within the city limits with access to park and recreation entities. This would include both temporary residents and permanent residents. Temporary residents would consist of students living within the city limits. This would include those in dormitories and fraternities/sororities pending they are within the city limits. Permanent residents would include permanent residents not limited to children, nursing home patients, and the homeless. Those in jail would be excluded since they currently do not have access to park and recreation entities.

**Exercise 2.3**

*A friend of yours is running for a place on the local school board, and you agree to help her by surveying local voters to learn which issues are most important and what they would like the school board to do. Can you get a list of registered voters who live in the school district? If so, does it contain mailing addresses? Telephone numbers? E-mail addresses? Apart from this, is there a directory of telephone numbers that might be usable for your purposes? A directory of mailing addresses? A directory of e-mail addresses?*

Depending on the location or district, registered voter lists from previous elections can be purchased from private companies. However, the district should have a list, whether it can be obtained is dependent on the district. It will most likely contain mailing addresses and possibly phone numbers. It is unlikely that email addresses will be listed. Any directory of email addresses is subject to coverage bias if it does not list a physical address within the district along with it.

**Appendix:**

HW02JackNelson.ipynb



