

ME001, MIIE01:

Information Systems Analysis and Design

Advanced Topics in Information Technology

Mini-project for optimal sample selection

It is known that the amount of data has been increasing tremendously in the last few years due to ease of access to internet, cheap or inexpensive mass storage devices, ease of transferring data through internet, communication lines and digital data are used in every walk of life. Nowadays, these big data have been used for data mining, knowledge discovery, machine learning, statistical learning, statistical analysis and experiments. In order to extract or discover useful data, information or knowledge from these big data, one of methods we usually could adopt is the sample selections.

In this mini-project, you are expected to extract a subset of samples from these big data. In order to extract this subset of data (samples), we have to make sure that the subset extracted or selected should be as fair and unbiased as possible and as optimal as possible. In the following one method is proposed.

Assume there are m samples ($45 \leq m \leq 54$), any n ($7 \leq n \leq 25$) out of m samples are selected. From one of these n samples, we randomly selected k ($4 \leq k \leq 7$) samples to form some groups. So there will be ${}_nC_k$ groups of k samples selected. Among these groups of selected k samples, we would like to optimize them by selecting ONLY some of them. The conditions that need to be fulfilled are listed as follows:

1. There are at least ONE group of k samples, in which s ($3 \leq s \leq 7$) samples have been selected from the j (where $s \leq j \leq k$) samples, i.e., when $j=4$, we have $s=3$ or 4 ; when $j=5$, we have $s=3, 4$ or 5 ; when $j=6$, we have $s=3, 4, 5$ or 6 ; and when $j=7$, we have $s=3, 4, 5, 6$ or 7 .

E.g. 1, when $m=45$, $n=7$ (assume we have chosen 7 samples, A, B, C, D, E, F, G and $k=6$, $s=5$, we could obtain the following minimum 6 groups of $k=6$ samples, which guarantee that at least ONE group of $k=6$ samples has the $s=5$ samples from $j=5$ out of $n=7$ samples, (i.e., ${}_nC_j$ samples).

- | | | |
|-----------------------|-----------------------|-----------------------|
| 1. A, B, C, D, E, G | 2. A, B, C, D, F, G | 3. A, B, C, E, F, G |
| 4. A, B, D, E, F, G | 5. A, C, D, E, F, G | 6. B, C, D, E, F, G |

E.g. 2, when $m=45$, $n=8$ (assume we have chosen 8 samples, A, B, C, D, E, F, G, H and $k=6$, $s=4$, we could obtain the following minimum 7 groups of $k=6$ samples, which guarantees that at least ONE group of $k=6$ samples has the $s=4$ samples from $j=4$ out of $n=8$ samples. (i.e., ${}_nC_j$ samples).

- | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|
| 1. A, B, C, D, G, H | 2. A, B, C, E, G, H | 3. A, B, C, F, G, H | |
| 4. A, B, D, E, F, G | 5. A, C, D, E, F, H | 6. B, C, D, E, F, H | 7. C, D, E, F, G, H |

E.g. 3, when $m=45$, $n=9$ (assume we have chosen 9 samples, $A, B, C, D, E, F, G, H, I$ and $k=6$, $s=4$, we could obtain the following minimum 12 groups of $k=6$ samples, which guarantees that at least ONE group of $k=6$ samples has the $s=4$ samples from $j=4$ out of $n=9$ samples. (i.e., ${}_nC_j$ samples).

- | | | | |
|-----------------------|------------------------|------------------------|------------------------|
| 1. A, B, C, D, E, I | 2. A, B, C, E, G, H | 3. A, B, C, F, H, I | 4. A, B, D, E, F, G |
| 5. A, B, D, G, H, I | 6. A, C, D, E, F, H | 7. A, C, D, F, G, I | 8. A, E, F, G, H, I |
| 9. B, C, D, F, G, H | 10. B, C, E, F, G, I | 11. B, D, E, F, H, I | 12. C, D, E, G, H, I |

E.g.4, when $m=45$, $n=8$ (assume we have chosen 8 samples, A, B, C, D, E, F, G, H and $k=6$, $s=5$, we could obtain the following minimum 4 groups of $k=6$ samples, which guarantees that at least ONE group of $k=6$ samples has the $s=5$ samples from $j=6$ out of $n=8$ samples. (i.e., ${}_nC_j$ samples).

- | | | | |
|-----------------------|-----------------------|-----------------------|-----------------------|
| 1. A, B, C, E, G, H | 2. A, B, D, F, G, H | 3. A, C, D, E, F, H | 4. B, C, D, E, F, G |
|-----------------------|-----------------------|-----------------------|-----------------------|

E.g. 5, when $m=45$, $n=9$ (assume we have chosen 9 samples, $A, B, C, D, E, F, G, H, I$ and $k=6$, $s=4$, we could

obtain the following minimum 3 groups of $k=6$ samples, which guarantees that at least ONE group of $k=6$ samples has the $s=4$ samples from $j=5$ out of $n=9$ samples. (i.e., ${}_nC_j$ samples).

1. A, B, D, F, G, H
2. A, C, E, G, H, I
3. B, C, D, E, F, I

E.g. 6, when $m=45$, $n=10$ (assume we have chosen 10 samples, $A, B, C, D, E, F, G, H, I, J$ and $k=6$, $s=4$, we could obtain the following minimum 3 groups of $k=6$ samples, which guarantees that at least ONE group of $k=6$ samples has the $s=4$ samples from $j=6$ out of $n=10$ samples. (i.e., ${}_nC_j$ samples).

1. A, B, E, G, I, J
2. A, C, E, G, H, J
3. B, C, D, F, H, I

E.g. 7, when $m=45$, $n=12$ (assume we have chosen 12 samples, $A, B, C, D, E, F, G, H, I, J, K, L$ and $k=6$, $s=4$, we could obtain the following minimum 6 groups of $k=6$ samples, which guarantees that at least ONE group of $k=6$ samples has the $s=4$ samples from $j=6$ out of $n=12$ samples. (i.e., ${}_nC_j$ samples).

1. A, B, D, G, K, L
2. A, C, D, H, J, L
3. A, D, E, F, I, L
4. B, C, G, H, J, K
5. B, E, F, G, I, K
6. C, E, F, H, I, J

2. A user friendly interface should be provided. A system name/title is given, e.g., “An Optimal Sample Selection System”.
3. The user can input the values for parameters m , n , k , j and s .
4. The user is asked to input the m numbers and n numbers out of m numbers.
5. Output groups of k samples to a DB file, e.g., 45-9-6-4-4- x for $m=45$, $n=9$, $k=6$, $j=s=4$ for the x th run.
6. Provide a mechanism to **DISPLAY** and **DELETE** the obtained groups of k samples onto the screen from a DB file, e.g., 45-9-6-4-4- x .
7. Students are required to form groups or teams yourself. Each group should have 3 or 4 students. You are advised to include in your team/group at least ONE student who knows how to do programming in MS ACCESS, version 2010.
8. Use numeral values, e.g., positive INTEGERS, 01,02,03,.....,54 instead of big letters A,B,C,D,E,F,....,Z for the m and n numbers.
9. Submit to me names of your team members NEXT WEEK.
10. A presentation and demonstration is a **MUST** in week 13.
11. Each group or team is required to have a **10-minute** presentation which includes the introduction, description of method(s) adopted, what have been achieved and/or not achieved in this project, and a demonstration of your project is a **MUST** in this presentation.
12. A clear, succinct, easy to understand **REPORT** of user manual/guide on how to **INSTALL and EXECUTE** your DEVELOPED system, **method(s)/methodology** you used in this project should be submitted in **hardcopy**.
13. You are required to submit a **USB** which contains your developed system, all your source files (codes), database files, DB files of k samples (outputs), and the **REPORT** mentioned in point 12.
14. Bonuses will be given to those group(s) that can allow users to select as many different parameters as possible for m , n , k , j and s , and could generate optimal or **near optimal** k samples. Furthermore, bonuses will be given to the developed system that could be executed in a short time, i.e., having good time complexity.
15. The deadline is **Week 13** in the presentation sessions. All teams must submit their projects in a **USB** and **hardcopy** of the **REPORT** in **Week 13**. Names, student numbers of your team members should be listed on your **REPORT**.