CAPTURE-RECAPTURE SAMPLING: A STUDENT PROJECT

Bruno C. de Sousa
Universidade do Minho - Campus de Azurém, Portugal
bruno@mct.uminho.pt

Capture-Recapture Sampling is used when one is interested in estimating the size N of a certain population. With the help of EXCEL, *we will illustrate how a student can estimate the size of a population using this technique. Different estimators for N will be considered, and the sampling distribution of the estimated values will be studied. Confidence intervals for N will be proposed and its interpretation will be presented. Some of the problems that we may encounter with this methodology will be briefly discussed, and a simulation in which the first captured individuals remain in a certain region of the population will be shown as an example of the effect that such behavior has on the estimated values.*

INTRODUCTION

Capture-Recapture Sampling is used to estimate the number of individuals in a population. We capture a first sample from the population and mark the individuals captured. If the individuals in a certain population are clearly identified, there is no need for any marking and consequently, we simply register these initially captured individuals. After an appropriate waiting time, a second sample from the population is selected independently from the initial sample. If the second sample is representative of the population, then the proportion of marked individuals in the second capture should be the same as the proportion of marked individuals in the population. From this assumption, we can estimate the number of individuals from a population. This procedure has been used not only to estimate the abundance of animals such as birds, fish, insects, and mice, among others, but also the number of minority individuals, such as the homeless in a city, for possible adjustments of undercounts in a census. In the following sections, we will present a simple procedure to simulate Capture-Recapture Sampling in *EXCEL*, followed by some possible procedures to estimate the size of a population. A brief discussion is given of some limitations of this method which may lead to biased results and is exemplified in a simulated situation.

ESTIMATING THE SIZE OF A POPULATION

Consider that we have a population of size N. We capture n_1 of these individuals and mark them in some way. After releasing them and waiting a proper amount of time, we capture a second sample of size n_2 from the population and count the marked individuals, calling it r. If the second sample is representative of the population under study, we expect the proportion of marked individuals in the second capture, r/n_2, to be the same as the proportion of marked individuals in the population, n_1/N. If we solve this equation for N we obtain the Peterson estimator $\tilde{N}=(n_1n_2)/r$. An estimator of the variance of $\tilde{N}$ is given by $v\tilde{a}r(\tilde{N})=[n_1n_2(n_1-r)(n_2-r)]/r^3$ (Sekar and Deming, 1949) and a simple approximate $(1-\alpha)100\%$ confidence interval is given by $\tilde{N}\pm z_{1-\alpha/2}\, sqrt(v\tilde{a}r(\tilde{N}))$, where the $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution.

COMPUTER SIMULATION

Consider that we have a population of size 250, i.e., N=250, and that we randomly mark 50 individuals on the first capture, n_1=50. For the second capture, we will randomly choose 25 individuals from the population, n_2=25, and count the ones that were marked in the first capture, r. In an *EXCEL* spreadsheet consider the following steps:

- In column A, generate N standard uniform random numbers.
- In column C, assign a serial number for each individual in the population (1 to N).
- In column D, copy the random numbers generated in column A, and sort columns C and D accordingly to column D in ascending order.
- In column E, mark the smaller 50 cases with an *M* as the marked individuals in the first capture.

- Sort all 3 columns (C, D and E) accordingly to column C in ascending order.
- For the second capture, copy columns C and E to columns G and I, respectively.
- Paste a copy of the values in column A into column H, and sort columns G, H and I accordingly to column H in ascending order.
- The first 25 individuals are the ones that belong to the second capture. Count how many of these 25 individuals are marked with an *M*.
- Compute an estimate *N* using the Peterson estimator.
- Repeat the process 100 times.

To illustrate this example, consider the following table where we exemplified how the above procedure is implemented in an *EXCEL* spreadsheet. The shaded areas represent the second captured individuals. In the first run of this experiment, we see that we have five individuals that were recaptured.

Known Population				First Estimation		
Random #	Serial No	Random #	1st Capture	Serial No	Random #	2nd Capture
0,34907	1	0,17910	M	214	0,00406	
0,11833	2	0,57852		7	0,00971	M
0,61670	3	0,70053		134	0,03132	
0,13488	4	0,18365	M	226	0,03625	M
0,12441	5	0,96857		209	0,04116	
0,50148	6	0,96994		132	0,04456	
0,71542	7	0,14066	M	98	0,04503	
0,97160	8	0,19156		50	0,04555	
0,78330	9	0,68201		138	0,04902	
0,37807	10	0,09489	M	147	0,04962	
0,77769	11	0,62419		79	0,06235	
0,58189	12	0,31447		203	0,06469	
0,04867	13	0,61458		219	0,06706	
0,52999	14	0,72963		97	0,06967	
0,43576	15	0,46698		217	0,07309	
0,21578	16	0,41363		118	0,07812	M
0,37822	17	0,10706	M	53	0,08360	
0,84007	18	0,45194		63	0,08432	
0,02525	19	0,67383		90	0,08698	
0,46762	20	0,86789		242	0,08772	
0,28276	21	0,59662		195	0,09347	
0,43000	22	0,23667		190	0,09494	M
0,73873	23	0,63166		114	0,09544	
0,64312	24	0,31157		110	0,09897	
0,09418	25	0,63464		201	0,11001	M
0,38389	26	0,83922		14	0,11257	
0,18737	27	0,89615		25	0,11601	
...	...	...	...	...	...	...

The next table represents the second and third repetition of the capture-recapture sampling method.

Second Estimation

Serial No	Random #	2nd Captured
37	0,00400	M
1	0,00831	M
166	0,00905	M
148	0,01016	
98	0,02520	
231	0,02529	M
64	0,03321	
118	0,03818	M
70	0,03867	
92	0,03988	
134	0,04356	
71	0,05199	M
240	0,05252	
142	0,05348	
11	0,05795	
67	0,05895	
220	0,06965	
141	0,06994	
127	0,07593	
249	0,09246	
181	0,09599	M
73	0,10809	
133	0,11250	
169	0,11390	
208	0,12635	
224	0,12665	M
100	0,12760	
...	...	...

Third Estimation

Serial No	Random #	2nd Captured
107	0,01320	
93	0,01885	
44	0,02406	
130	0,03007	
86	0,03023	M
70	0,03675	
208	0,04360	
202	0,05227	
1	0,07098	M
81	0,07547	
94	0,07811	
15	0,07905	
62	0,08656	M
119	0,10677	M
220	0,10849	
232	0,10894	
20	0,10974	
134	0,11452	
142	0,11690	
184	0,12370	
76	0,13014	M
31	0,13097	M
210	0,13321	
174	0,13428	M
241	0,13513	
79	0,13742	
123	0,13793	M
...	...	...

The following table shows the estimated values obtained for 4 repetitions of the procedure.

	SAMPLE						
Pop. size	First marked	Second Capture	1st Recapture	2nd Recapture	3rd Recapture	4th Recapture	...
250	50						
		25	5	7	7	4	...

Estimate of N	250	178,6	178,6	312,5	...

The students will immediately realize that the values obtained in each run vary a lot, with some of these values being very close to or even the same as the true value of N=250. Were some of the runs done incorrectly? This is just one of the questions that might come up naturally in after a brief analysis of the results. At this point, a histogram of the estimated values will be drawn and the concept of the sampling distribution of $\tilde{N}$ is introduced. Once the concept of the sampling distribution is understood, students might suggest the use of different statistics to estimate the size of the population N, for example, the average of the estimated values or the mean excluding the larger and smaller values, or even the median of the estimated values.

We believe that at this time in the project we also can introduce very naturally the concepts of a Confidence Interval for N, calculating them based on confidence intervals for the proportion of marked individuals in the population, n_1/N. Under certain conditions, we have the (1-α)100% confidence interval for n_1/N given by $r/n_2 \pm z_{1-\alpha/2}$ $sqrt([r/n_2(1-\ r/n_2)]/n_2) = (a,b)$.Therefore, the (1-$\alpha$)100% confidence interval for N is given by $(n_1/a,\ n_1/b)$. By calculating the confidence intervals for N, students can more clearly see its interpretation and thus avoid misleading conclusions of Confidence Intervals in future studies.

WHERE ARE THE PROBLEMS?

Some of the most common problems that lead to biased results in the estimation of the size of a population in Capture-Recapture Sampling result from what is known as *trap-shy* or *trap-happy* behavior. For example, when talking about populations of animals, some animals might have a different behavior after the first capture. Some might get used to human contact and therefore the probability of their capture in the second sample his higher then the rest of the population. To illustrate this behavior, we will repeat the simulation presented above where the second captures are obtained with the help of generated Chi-square distribution values with 20 degrees of freedom instead of the standard uniform distribution. Since the new distribution is skewed to the right we expect that the individuals with lower Chi-square values will have a higher probability of being selected in the second sample drawn from the population. This is the *trap-happy* behavior which usually leads to an underestimation of *N*.

To illustrate this example, consider the following table where we present the results of the first 4 recaptures.

Pop. size	SAMPLE First marked	 Second Capture	1st Recapture	2nd Recapture	3rd Recapture	4th Recapture	...
250	50						
		25	6	6	3	8	...

Estimate of N	208,3	208,3	416,7	156,3	...

As predicted, most of the estimated values are lower than the true size of the population, *N*=250. As in the previous section, we can perform a similar analysis of the results and study the effect of the *trap-happy* behavior on the Capture-Recapture Sampling method.

CONCLUSION

Capture-Recapture Sampling is used in many different situations in real life, which leads us to consider many different objectives where a variety of problems may surface. With this Student Project, we hope to transmit these ideas to the students and engage them, thus sparking lively discussions on the subject. The ideas of Sampling Distribution and Confidence Intervals are quite difficult for students to grasp. Presenting this project in *EXCEL* allows the students to put their knowledge into practice and, on their own, apply it to other situations, thus enabling each student to actually experiment with these concepts. Future work can be directed toward the area of trying to answer questions such as, how the sample sizes affect the estimation of the size of the population *N*, or how we can improve the estimation process of *N*, or even how Bootstrap Methods help in the Capture-Recapture Sampling. These are definitely complex questions to address, but this can become a challenge for students and will allow for the introduction of many current subjects in statistics.

ACKNOWLEGMENTS

I wish to thank Alfredo Aliaga for his insights into the problem and for our discussions on the subject.

REFERENCES

Aliaga, M. and Gunderson, B. (2002). *Interactive Statistics* (2nd edition). New York: John Wiley and Sons, Inc.

Landwehr, J. M., Swift, J. and Watkins, A. E. (1987). *Exploring Surveys and Information from Samples* (1st edition). New Jersey: Dale Seymour Publications.

Sekar, C. C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.

Thompson, S. K. (2002). *Sampling* (2nd edition). New York: John Wiley and Sons, Inc.