A Beginner's Resource to Capture-Recapture Estimation

Kathryn McGinnis, Prof. James Broda

5/18/2021

An Alternative Estimation Method

Increased rates of homelessness, especially among youth, are a growing concern for American cities. Improper housing, unsafe family situations, and addiction are driving more people each year to be temporarily unhoused. While the need for effective housing policy is great, governments and community agencies must also have proper evaluation programs to monitor the impact of these policies.

The Department of Housing and Urban Development (HUD) conducts an annual Point-In-Time (PIT) survey to estimate the number of homeless youth in a given area. The data is an important metric for evaluating a community's need, but it is a limited picture of people experiencing homelessness at any point in time. The surveyor only counts the number of people *observed*, including people in temporary shelters and those sheltering outside. It does not provide an estimate for the number of unhoused people the surveyor did not come into contact with. This is known as the *unobserved* population.

People experiencing homelessness may be more reluctant to engage with social programs due to a history of incarceration or general distrust. Their elusiveness is one of the difficulties inherent in estimating population size. In addition, an unhoused person may not be unhoused each night. Many people couch-surf with friends or live in cars for a short period of time. A static survey count of the number of people experiencing homelessness on a given night will not reflect the changing nature of the population.

The Capture-Recapture method emerged as a technique to estimate the size of endangered animal populations. Rather than attempt to observe each member of a population, it counts the number of times a researcher comes into contact with the same individual. These *capture histories* are then used to estimate the size of the observed population (the number of unique individuals the researcher observed) and the unobserved population (the members of the population the researcher did not see).

Data Collection

Before a community agency can begin a Capture-Recapture experiment, it must decide what the design of the experiment will look like.

1. Are people moving in and out of the population during the study period?
2. How many days are you able to collect data?
3. Are you interested in additional demographic characteristics of the population such as age, gender, or education level?

The answers to these questions will determine the specific type of Capture-Recapture experiment you can conduct. In this document I will show you how to design and analyze a simple closed-population model with multiple capture histories. While simpler models are easier to estimate, community agencies should consider whether the model's assumptions reflect the situation in the field. For example, in a closed population, no one enters or exits the population during the study period. This assumption is impractical in homeless populations if the data collection period spans multiple months, but may be true over a span of a few days or weeks. The limiting assumptions an agency chooses to include in its model will affect the interpretation of results at the end.

To estimate a closed population Capture-Recapture experiment, a researcher must record when a member of the population was observed and whether or not they have been seen before throughout the data collection period. The actual data can take many forms, including cross referencing unique lists, recording identifying characteristics of population members, or asking people to recall images they have seen in the past.

Example Data Collection: Allegheny County, PA Lesbian Population

This data was taken from a 2002 study completed by Aaron, Chang, Markovic, and LaPorte to estimate the number of lesbians in a Pennsylvania county. The researchers analyzed four mailing lists from LGBTQ community centers and organizations throughout the county, noting the number of women that appear on multiple lists. In this design, each list will serve as a separate observation period.

The data for this experiment are available to the public and preloaded into the Rcapture software; however, the software does not show how the data was manipulated into a usable form for estimation. I will show one method to do this from an artificial data set.

Working with Rcapture

To get started, download the computer program "R" and "R Studio" onto your computer. The Capture-Recapture estimation program is a free package you can add to your standard program. Once you open R,

1. Select Packages/Tools at the top of the screen
2. Click "Install Packages"
3. Select USA
4. Find and select "Rcapture"

Viewing the Data

Now we are going to upload an example csv file that shows what the raw data in this lesbian study may have looked like.

```
library(Rcapture)
df<-read.csv("Example Dataset.csv")
df<- df[,-4]
df
```

```
##               X1            X2            X3
## 1  G. Washington G. Washington       B. Ross
## 2    B. Franklin  T. Jefferson   B. Tarleton
## 3    A. Hamilton     B. Arnold     F. Marion
## 4   T. Jefferson      S. Adams       M. Hays
## 5      P. Revere  J. Dickinson      S. Adams
## 6      B. Arnold   J. Trumbull     N. Greene
## 7       S. Adams      J. Brant   J. Trumbull
## 8      N. Greene      H. Gates G. Washington
## 9       P. Henry   J. Sullivan    J. Hancock
## 10       N. Hale    B. Lincoln
## 11    C. Attucks
## 12    J. Hancock
## 13       H. Knox
## 14     D. Morgan
```

Creating a Capture History

The goal is to create a capture history matrix that the Rcapture program will use to estimate population size. A capture history is a personalized record for each member of a population. It indicates when an individual

is sighted throughout the observation period.

First, aggregate each individual list of names to create a master list.

```
#Label each list
a<-df[,1]
b<-df[,2]
c<- df[,3]

#Create one master list
total<-c(a,b,c)
Total<-total[total !=""]
Total
```

```
##  [1] "G. Washington" "B. Franklin"   "A. Hamilton"   "T. Jefferson"
##  [5] "P. Revere"     "B. Arnold"     "S. Adams"      "N. Greene"
##  [9] "P. Henry"      "N. Hale"       "C. Attucks"    "J. Hancock"
## [13] "H. Knox"       "D. Morgan"     "G. Washington" "T. Jefferson"
## [17] "B. Arnold"     "S. Adams"      "J. Dickinson"  "J. Trumbull"
## [21] "J. Brant"      "H. Gates"      "J. Sullivan"   "B. Lincoln"
## [25] "B. Ross"       "B. Tarleton"   "F. Marion"     "M. Hays"
## [29] "S. Adams"      "N. Greene"     "J. Trumbull"   "G. Washington"
## [33] "J. Hancock"
```

```
#Notice that this list contains repeated names
#Some people were sighted multiple times during the collection period.
```

Next, identify the number of unique names that appear on each list.

```
#Identify the unique names in the master list
Unique<-unique(Total)
Unique
```

```
##  [1] "G. Washington" "B. Franklin"   "A. Hamilton"   "T. Jefferson"
##  [5] "P. Revere"     "B. Arnold"     "S. Adams"      "N. Greene"
##  [9] "P. Henry"      "N. Hale"       "C. Attucks"    "J. Hancock"
## [13] "H. Knox"       "D. Morgan"     "J. Dickinson"  "J. Trumbull"
## [17] "J. Brant"      "H. Gates"      "J. Sullivan"   "B. Lincoln"
## [21] "B. Ross"       "B. Tarleton"   "F. Marion"     "M. Hays"
```

```
#24 unique names appear on the three lists
```

Now that the program has identified the name of each unique individual, it can search each list for a specific name. If a name appears on a given list, the program will show a "1". If the name is not on the list, the program will show a "0".

```
#Create an empty matrix to store the capture histories

History<-NULL
x1<-rep(0, length(Unique))
x2<-rep(0, length(Unique))
x3<- rep(0, length(Unique))
History<- cbind(x1, x2, x3)


#Search each list in the dataset to see if it contains the unique name
#Example shown for first five unique names
```

```
x<-NULL
for (k in 1:3){x[k]<- sum(df[,k]== 'G. Washington')}
n1<-as.numeric(x)
n1
```

```
## [1] 1 1 1
```

```
#This capture history implies that G. Washington appeared on each list

for (k in 1:3){x[k]<- sum(df[,k]== 'B. Franklin')}
n2<-as.numeric(x)
n2
```

```
## [1] 1 0 0
```

```
#This capture history shows that B. Franklin only appeared on the first list

for (k in 1:3){x[k]<- sum(df[,k]== 'A. Hamilton')}
n3<-as.numeric(x)
n3
```

```
## [1] 1 0 0
```

```
for (k in 1:3){x[k]<- sum(df[,k]== 'T. Jefferson')}
n4<-as.numeric(x)
n4
```

```
## [1] 1 1 0
```

```
#T. Jefferson appeared on the first and second list

for (k in 1:3){x[k]<- sum(df[,k]== 'P. Revere')}
n5<-as.numeric(x)
n5
```

```
## [1] 1 0 0
```

```
All<- rbind(n1, n2, n3, n4, n5)
All
```

```
##    [,1] [,2] [,3]
## n1    1    1    1
## n2    1    0    0
## n3    1    0    0
## n4    1    1    0
## n5    1    0    0
```

```
CaptureHistory<- NULL
CaptureHistory<- cbind(Unique[1:5], All)
CaptureHistory
```

```
##    [,1]            [,2] [,3] [,4]
## n1 "G. Washington" "1"  "1"  "1"
## n2 "B. Franklin"   "1"  "0"  "0"
## n3 "A. Hamilton"   "1"  "0"  "0"
## n4 "T. Jefferson"  "1"  "1"  "0"
## n5 "P. Revere"     "1"  "0"  "0"
```

A capture history matrix is a helpful tool for organizing and analyzing data. The example shown above is one way to present a capture history matrix. The lesbian data set that we will analyze for the remainder of

this paper uses a capture history format very similar to the one above.

```
#Examine the capture history matrix for the lesbian data set
lesbian
```

```
##       A B C D freq
##  [1,] 1 1 1 1   23
##  [2,] 1 1 1 0   27
##  [3,] 1 1 0 1   48
##  [4,] 1 1 0 0  104
##  [5,] 1 0 1 1   19
##  [6,] 1 0 1 0   44
##  [7,] 1 0 0 1  143
##  [8,] 1 0 0 0  589
##  [9,] 0 1 1 1   20
## [10,] 0 1 1 0   64
## [11,] 0 1 0 1   53
## [12,] 0 1 0 0  534
## [13,] 0 0 1 1   28
## [14,] 0 0 1 0  281
## [15,] 0 0 0 1  208
```

```
#This matrix shows each unique capture history
#It also shows the number of times each capture history appears
```

Identifying Hetereogeneity

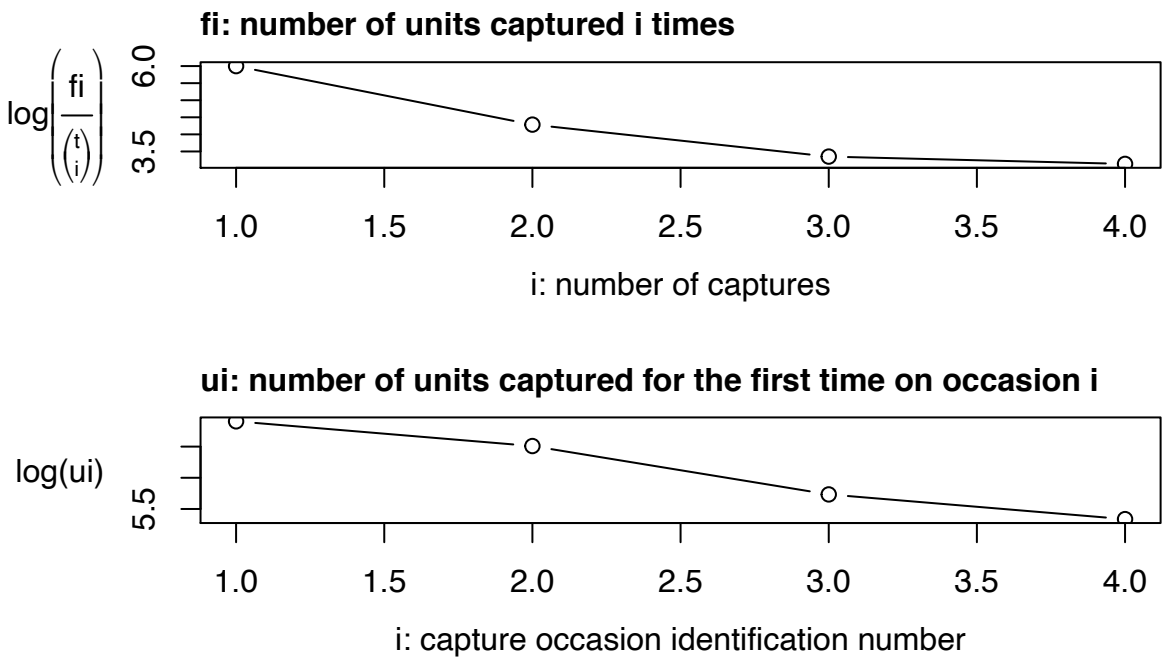
In simple Capture-Recapture designs, researchers may assume that all members of the population have an equal probability of observation; however, this assumption is difficult to meet. Natural heterogeneity can occur in a population for a variety of reasons, including differences in age, gender, or behavior. If a portion of the population is more difficult to observe than others, the population estimate may be biased.

We can evaluate our dataset for heterogeneity by creating a plot.

```
desc<-descriptive(lesbian, dfreq=TRUE)
desc
```

```
##
## Number of captured units: 2185
##
## Frequency statistics:
##        fi    ui    vi    ni
## i = 1  1612   997   589   997
## i = 2   436   671   638   873
## i = 3   114   309   416   506
## i = 4    23   208   542   542
## fi: number of units captured i times
## ui: number of units captured for the first time on occasion i
## vi: number of units captured for the last time on occasion i
## ni: number of units captured on occasion i
```

```
plot(desc)
```



If each member of the population had an equal probability of observation, these graphs would appear linear. As you can see, the slopes are not quite constant. We correct for the presence of heterogeneity by using a Capture-Recapture model that estimates different observation probabilities for each member of the population.

Running the Calculation

Now it is time to estimate the size of the population.

```
closedp(lesbian, dfreq=TRUE)
```

```
##
## Number of captured units: 2185
##
## Abundance estimations and model fits:
##                abundance stderr deviance df     AIC     BIC infoFit
## M0                3829.2   99.4  477.249 13 573.601 584.980      OK
## Mt                3738.5   95.2  174.809 10 277.162 305.609      OK
## Mh Chao (LB)      4420.0  161.5  408.316 11 508.669 531.426      OK
## Mh Poisson2       5027.0  247.5  411.124 12 509.477 526.545      OK
## Mh Darroch        6955.3  649.1  408.337 12 506.690 523.758      OK
## Mh Gamma3.5      10001.8 1485.3  408.547 12 506.900 523.968      OK
## Mth Chao (LB)     4321.4  155.1   98.108  8 204.461 244.286      OK
## Mth Poisson2      4942.2  240.3  101.296  9 205.648 239.784      OK
## Mth Darroch       6963.7  650.1   98.134  9 202.487 236.623      OK
## Mth Gamma3.5     10259.6 1535.5   98.362  9 202.714 236.850      OK
## Mb                2470.2   33.0  191.372 12 289.724 306.792      OK
## Mbh               2402.8   43.4  184.862 11 285.214 307.972      OK
```

The code closedp shows estimation results for various Capture-Recapture models. The models with a subscript h correct for heterogeneity in the population while the subscript t adjusts for temporal differences between lists. As you can see from the table, depending on the type of model used, the population estimate can

vary wildly. Since the actual number of lesbians in Allegheny County is unknown, researchers cannot sort model types by their accuracy. Instead, consider the Akaike Information Criteria (AIC) reported in the table. The AIC is a measure of the information lost during estimation. It can be used to determine which model synthesizes information more efficiently than others. According to this table, the Mth Darroch model is the best to use. We do not know if this model is also the most accurate.

Interpeting Results

The Capture-Recapture model estimates 6,964 lesbians residing in Allegheny County, PA with a deviance of 98 people. It is necessary to report the deviance because statistical methods only estimate the most likely population estimate given a particular set of data. There is always a chance that the true value of a population lies in the tail ends of a distribution.

While the true value of the lesbian population in Allegheny County cannot be determined from statistical estimation methods, there is still useful information that can be gained from Capture-Recapture estimates. In the original data, 2,185 unique women appeared on the mailing lists, with nearly 75% of women appearing on only one list. The population estimate predicts 4,779 additional lesbians residing in the county that did not appear on any lists. This estimate for the unobserved population is a key feature of Capture-Recapture estimation. Community agencies may use this analysis to determine how many people their services could reach in the future.

Estimating the unobserved population is a special feature of the Capture-Recapture model; however, it may not encompass the entire population. In this example, the population estimate is actually measuring the number of lesbian women in Allegheny County that may engage with LGBTQ community organizations. It does not provide insight into the population size of all lesbians in the community, including people that do not engage in public organizations or people that choose to not be out publicly.

Conclusion

Capture-Recapture estimation can be an invaluable source to community agencies. Understanding the number of people a program has the ability to reach is crucial to future planning. Estimates of the unobserved population may also lead to better informed funding applications, especially in areas where a population is particularly difficult to observe. This handout should serve as basic instructions for completing a Capture-Recapture experiment. Further research into Capture-Recapture estimation in human populations is ongoing around the world and I hope it will prove to be a cost-effective and reliable method to measure the size of vulnerable populations.