18

Capture–Recapture Methods for Estimating the Size of a Population: Dealing with Variable Capture Probabilities

Louis-Paul Rivest and Sophie Baillargeon

Université Laval, Québec, QC

18.1 Estimating Abundance with Marked Units

Demographers predict that the Canadian population will reach the 35 million mark in 2013; this represents a 4% increase over a 5-year period. This is a projection of the results of the 2011 census that shows that the Canadian population is growing at a steady pace. In a similar way, the size of an animal population is a basic demographic characteristic that reflects its general well-being. Consider, for instance, the George River caribou herd in Northern Québec and Labrador. Thirty years ago, it was among the most expansive groupings of large mammals on the planet, with more than 600,000 animals. Now the herd has dwindled and is monitored closely. As this is a migratory animal, estimating the herd's size is a challenge. The current method uses animals marked with GPS collars to locate and photograph the large groups that form during the short nordic summer. A complete enumeration is not possible and final estimates are calculated by "entering the number of caribou found by biologists into an algorithm that determines the herd's size" as explained in a recent article in a Montréal newspaper. The statistical model behind this algorithm is detailed in Rivest et al. (1998) and several ways to account for missed animals are available. The estimate for the summer 2012 inventory is 27,600 caribou and protective measures, such as reducing hunting permits, have been taken.

The caribou inventory provides an example of an ingenious method developed by statisticians and biologists for estimating the size of a wildlife population using marked animals. According to Lecren (1965), marked animals have been used in ecology since the start of the 20th century. This con-

sists of tagging animals with a unique identifier that allows for identification on subsequent captures or, more generally, encounters. Nowadays, advances in molecular biology broaden the scope of applications of capture-recapture techniques as they permit capturing and identifying animals through DNA collected in feces or hair snags. One distinguishes investigations of open populations, whose main objective is the estimation of survival rates, from those of closed populations that focus on population sizes; see Williams et al. (2002). This work is concerned with the latter; it assumes that the number of deaths and of births is negligible during the survey period and that the size of the population is, for all practical purposes, constant. Capture-recapture is the main method for estimating the abundance of animals not easily visible on the ground, such as the harvest mouse population discussed in this work. The previous chapter, by Cowen, Challenger, and Schwarz, describes some other applications of capture-recapture techniques.

Let N be the unknown size of the population of interest. Its estimation by capture-recapture typically involves the capture and the recapture of animals over a short period. The study discussed in the next sections has $t = 14$ weekly capture occasions. At the start, a grid of live traps is laid out in the population habitat. All the traps are visited on each capture occasion. When an animal is captured for the first time, it is marked with a unique identifier and released. Thus, when it is recaptured, an animal bearing a mark can be identified and linked with its previous captures. At the end of the study, each animal has its own capture history; it consists of 14 entries, equal to either 0 or 1, where a 1 in position i means that the animal has been captured at occasion i. Covariates, such as sex and body weight, are sometimes recorded as they might be related to the number of times that an animal is captured. The probability of being captured is often the same for all capture occasions. In such a case, the complete capture history is not useful, only the total number of captures is reported. The dataset is then given by $\{(c_j, x_j) : j = 1, \ldots, n\}$ where n is the number of animals caught at least once, c_j is the number of captures for animal j, and x_j represents covariates measured on that animal. Animals with the capture history $(0, \ldots, 0)$ are never caught and their number is unknown; the goal of the analysis is to estimate that number.

In ecology, species richness can be estimated using standard capture-recapture techniques for closed populations. In this context, a unit is a species and a capture occasion is an observation station. A unit is observed at a station if it is seen or heard at that station; the goal is to estimate the total number of species in an area, including those that went undetected. In this example, the species detected at an observation station are a partial or incomplete list of all the species in the area. Viewing a capture occasion as an incomplete list of units makes capture-recapture techniques also applicable to elusive human populations. For instance, the total number of illegal drug users in a city can be estimated through partial lists of users provided by police stations, drug treatment services and community centers. In a similar way, partial lists of patients coming from hospital records, drug prescriptions and doctors' offices provide

the basis for estimating the number of people suffering from a particular disease in an administrative region. When applied to administrative records, capture-recapture methods are called multiple system estimation; linking the records from various lists together in order to create capture histories raises interesting statistical issues. Nowadays the wide availability of administrative data broadens the application areas of capture-recapture techniques. For example, multiple system estimation is used extensively by the Human Rights Data Analysis Group to evaluate human rights violations, such as killings, in troubled areas; see Harrison (2012) for a discussion of the methodological challenges involved.

Capture-recapture also applies in investigations without clearly defined capture occasions. A famous example, due to McKendrick (1926), attempts to estimate the number of households affected by cholera in an Indian village from a list containing the name and the address of persons suffering from cholera. The addresses identify households and the list allows to determine the number of members affected, or "captured," by cholera in each one. In some disease free households the cholera might be latent and appear later; the goal is to estimate the number of such households. In this example a household is captured when at least one of its member is diagnosed with cholera. There are no well defined capture occasions and "continuous captures" describes such a dataset.

Capture-recapture models attempt to extrapolate, to the whole population, observations made on captured units. This works as long as the units captured are representative of the whole population. However, consider a simple situation where the population consists of two groups, say 1 and 2. If the units in Group 1 are much easier to capture than those in Group 2, then capture-recapture estimates are derived mostly from Group 1. They do not apply to the whole population unless the units caught in Group 2 are given more weight to account for their underrepresentation. Indeed if the "unit level covariate" group could be recorded, then the solution would be to estimate the size of the two groups separately. When some units are easier to catch than others, the capture probabilities are said to be variable or heterogeneous. This is quite common. When studying animals, the size of the animal's home range and its weight might be positively associated with its capture probability while the detection of a species depends on the importance of its habitat in the study area. In epidemiological applications, the severity of an illness is positively correlated with the probability of appearing in a partial list of patients. Such heterogeneity in capture probability makes the estimation of N difficult as the data might not be representative of the whole population. The solution is to measure "unit level covariates," such as the group identifier alluded to earlier, and to use them to model the capture probabilities before estimating N. Even when this is done, the covariates might fail to explain all of the heterogeneity and the estimate of N might be biased. The goal of this chapter is to review the statistical methods available to detect heterogeneity in capture probabilities and to account for it when estimating N. These meth-

ods are presented through the analysis of two datasets that are introduced in the next section.

18.2 Datasets

The first dataset considered here has been obtained from a study of the harvest mouse (*Micromys minutus*) conducted at the Wulin Recreation Area in Shei-Pa National Park, Taiwan, in the summer of 2008. Such small mammal populations are investigated in ecology since they give handy indicators of the impact of human interventions, such as wood cutting and reforestation, on wildlife habitats. The data are presented in Stoklosa and Huggins (2012) and Stoklosa et al. (2011); there are $t = 14$ capture occasions. A total of $n = 142$ different mice have been captured over the 14 occasions. The number of captures per mouse ranged from 1 to 10, and the body weight of a mouse was found to be related to the capture probabilities. Other covariates such as sex and hindfoot measurement were available; however, they were unrelated to the capture probabilities and are not considered here. This dataset is available in the R package `PL.popN` (Stoklosa, 2012) in the form $\{(c_j, x_j) : j = 1, \ldots, n\}$, where $c_j \geq 1$ is the number of times that mouse j is captured and x_j is the weight of mouse j. Leaving the weight variable aside, the data can be presented in aggregated form as $f_1 = 68$, $f_2 = 29$, $f_3 = 16$, $f_4 = 8$, $f_5 = 10$, $f_6 = f_7 = 4$, $f_8 = 2$, $f_{10} = 1$, and $f_9 = f_{11} = f_{12} = f_{13} = f_{14} = 0$, where f_i is number of animals captured i times, that is the number of occurrences of $c_j = i$ among the 142 mice that have been captured.

The second dataset is an instance of continuous captures. It concerns illegal immigrants in four large cities in the Netherlands over a one-year period. These immigrants cannot be effectively expelled and are therefore liable to be captured several times by the police; see van der Heijden et al. (2003) for a detailed presentation. We use the portion of the data available in Supplement C to Böhning and van der Heijden (2009). It features $n = 799$ illegal immigrants apprehended once or more. The covariates possibly related to their detection are their sex, a dichotomized age (above or below 40 years old), their country of origin with five possible values, and their reason for being arrested. The maximum number of encounters was 6 and the aggregated data gives $f_1 = 686$, $f_2 = 90$, $f_3 = 18$, $f_4 = 3$, $f_5 = f_6 = 1$.

In these two examples, the goal is to estimate the total population size N that includes the units that were missed. In the mouse example, N may provide an evaluation of the quality of the site's habitat while in the illegal immigrant example it gives an indication of the extent of this social problem in the Netherlands. The plots and the calculations presented in Section 18.3 have been carried using `Rcapture`, an R package for capture-recapture models presented in Baillargeon and Rivest (2007).

18.3 Estimation of Population Size Using Aggregated Data

This section reviews methods to estimate the population size N without using the unit level covariates. The first comprehensive presentation of models and methods for doing so is the monograph *Statistical Inference from Capture Data for Closed Animal Populations* by Otis et al. (1978) along with `CAPTURE`, its companion software. It distinguishes three types of effects that need to be addressed when estimating a population size, namely a time effect, a behavioral effect, and an heterogeneity effect. This work focuses on the latter, as it is also a concern in multiple system estimation of human populations and in studies featuring continuous captures.

Otis et al. (1978) used the acronym M_h to denote the heterogeneity model while M_0 represents the homogeneity model. Only estimates of N derived from statistical models for the data are considered. The jackknife estimator, used by Otis et al. (1978), does not fall in this category as it does not rely on a particular model. The modern statistical methodology on the estimation of a population size emphasizes fitting a model as an important intermediate step in the analysis. Models providing a good fit to the data are expected to give reliable estimates of population sizes. We start with the homogeneity model M_0 in the next section.

18.3.1 Homogeneity Model

For a capture-recapture dataset featuring t capture occasions, the homogeneity model postulates that there is a positive probability p for a unit to be captured at a given occasion and that it is the same for all units and all occasions. The number of times C that a unit is captured is a random variable with what is known as a binomial distribution with parameters p and t for all units in the population. The probability that C is equal to i is then

$$\Pr(C = i) = \binom{t}{i} p^i (1-p)^{t-i}, \quad i \in \{0, \ldots, t\}.$$

Under this model the probability that a unit is not captured is $(1-p)^t$, thus on average $N(1-p)^t$ will be missed in the study, and an estimate of N is given by n plus an estimate of $N(1-p)^t$.

To obtain a numerical value for $N(1-p)^t$, we fit the homogeneity model to the data $\{f_i : \ i = 1, \ldots, t\}$. Given that C has a binomial distribution, the expectation (or predicted value) for f_i is given by

$$\mathrm{E}(f_i) = N \binom{t}{i} p^i (1-p)^{t-i}$$

or

$$\mathrm{E}(f_i) = E_i = \binom{t}{i} e^{\gamma + i\beta}, \quad i \in \{1, \ldots, t\}, \tag{18.1}$$

where $\gamma = \log\{N(1-p)^t\}$ is the logarithm of the predicted frequency for the missed units and $\beta = \log\{p/(1-p)\}$ is called the logit of the capture probability p. For fixed values of (γ, β) the discrepancy between the observed values f_i and the predicted values E_i is measured by what statisticians term the deviance, given by

$$G^2 = 2\sum_{i=1}^{t}\{f_i \log(f_i/E_i) - (f_i - E_i)\}. \tag{18.2}$$

Note that the value of G^2 is small when the E_i closely approximate the observed frequencies f_i. The values of (γ, β) for which G^2 is minimum are called the maximum likelihood estimates of the parameters and are denoted $(\hat{\gamma}, \hat{\beta})$. These are the values that will be used to estimate N,

$$\hat{N} = n + e^{\hat{\gamma}}. \tag{18.3}$$

In statistics, model (18.1) is called a generalized linear model. Besides providing the point estimates $(\hat{\gamma}, \hat{\beta})$, it also gives variance estimates $(v(\hat{\gamma}), v(\hat{\beta}))$ as measures of uncertainty. The sampling variance of $\hat{N}$ can then be estimated by

$$v(\hat{N}) = e^{\hat{\gamma}} + v(\hat{\gamma})e^{2\hat{\gamma}}; \tag{18.4}$$

see Rivest and Lévesque (2001) for more discussion. In Tables 18.1–18.3, the standard error (SE) is the square root of the variance estimate $v(\hat{N})$ in (18.4) and the coefficient of variation (CV), defined as the ratio $SE/\hat{N}$, gives a standardized measure of precision. Roughly, estimates with a CV smaller than 5% are considered as being accurate while CVs larger than 20% are considered imprecise.

For continuous captures, the distribution for the number of captures under the homogeneity model is the Poisson distribution,

$$\Pr(C = i) = \frac{1}{i!}\,\lambda^i\, e^{-\lambda}, \quad i \in \{0, 1, \ldots\},$$

where λ, the "encounter rate," is the average number of detections per unit time. For the illegal immigrant data, λ represents the average number of captures per year. In the homogeneity model, λ is assumed to be the same for all units in the population. One has

$$\mathrm{E}(f_i) = E_i = \frac{1}{i!}\, e^{\gamma + i\beta}, \quad i \in \{1, \ldots, t_{\max}\}, \tag{18.5}$$

where $t_{\max}$ is the largest i for which $f_i > 0$, $\gamma = \log(Ne^{-\lambda})$ and $\beta = \log\lambda$. Once the generalized linear model for (18.5) is fitted, the estimates of N and of its variance are calculated using Equations (18.3) and (18.4).

TABLE 18.1: Homogeneity and lower bound (LB) models fitted to the harvest mouse and to the illegal immigrant data.

	Harvest Mouse Data			Illegal Immigrant Data		
	M_0	LB	M_0 ($t_0 = 3$)	M_0	LB	M_0 ($t_0 = 3$)
$\hat{N}$	158	216	195	2,765	3,413	3,002
SE	5.0	24.2	14.4	205.6	344.1	243.8
CV	3.1%	11.2%	7.4%	7.4%	10.1%	8.1%
G^2	81.9	—	2.6	22.1	—	6.2
df	12	—	1	4	—	1

One method used to assess whether a model fits the dataset $\{f_i\}$ is to compare the deviance statistic G^2 for the fitted model with its degrees of freedom. The number of degrees of freedom is t (or $t_{\max}$), the number of f_i-values, minus the number of parameters in the model. A model is said to fit well if its deviance is small, in particular smaller than its degrees of freedom. For model M_0, the deviance has $t-2$ degrees of freedom. Table 18.1 gives the deviances of the homogeneity model obtained for the two datasets of Section 18.2. The two deviances are much larger than their degrees of freedom indicating a bad fit. Thus the M_0 estimates of N for the two datasets, 158 and 2,765, should be discarded.

18.3.2 A Probability Plot and a Lower Bound Estimate

Under model (18.1), $\log\{\mathrm{E}(f_i)/\binom{t}{i}\}$ is a linear function of i. Thus a simple binomial probability plot to investigate this assumption contains the points $(i, \log\{f_i/\binom{t}{i}\})$ for all i for which $f_i > 0$. If the plot is approximately linear, then the homogeneity model fits well. For continuous captures, a Poisson probability plot considers the points $(i, \log(i!f_i))$ for the positive f_i's. It should be linear, up to sampling errors, if the homogeneity model fits well; see Hoaglin (1980) for more discussion. The probability plots for the two datasets are shown in Figure 18.1 together with lines, $\hat{\gamma} + i\hat{\beta}$, representing the fitted homogeneity models. The two plots are not linear and the lines for the homogeneity model do not represent the points with large i-values well. This agrees with the conclusion of Section 18.3.1 that the homogeneity model is inadequate.

The homogeneity model fails when the capture probability p changes from one unit to the next. This feature is included in the statistical model by considering p as a random variable rather than a fixed parameter. In this case $\Pr(C = i)$ is proportional to

$$\binom{t}{i}\mathrm{E}\{p^i(1-p)^{t-i}\},$$

where $\mathrm{E}(\cdot)$ gives an average value over the units in the population. One can

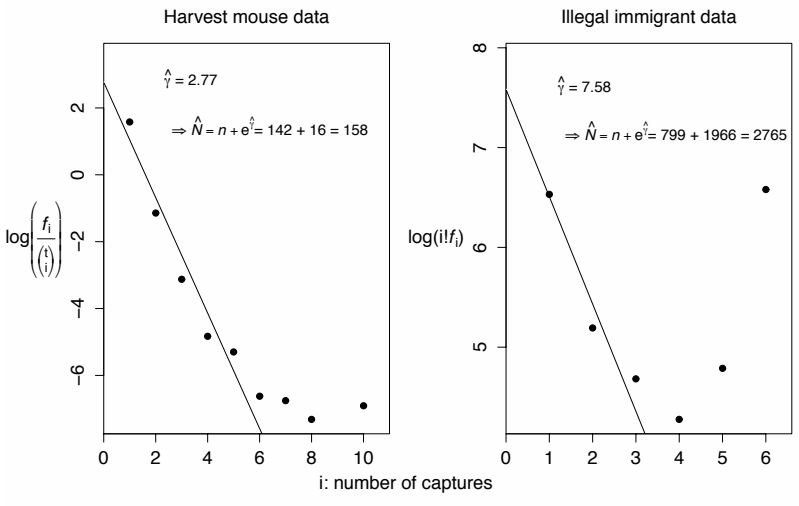

FIGURE 18.1: Probability plots for the two datasets with graphical representations of the fit of the homogeneity model.

show that when p is random $\log\{\mathrm{E}(f_i)/\binom{t}{i}\}$ is a convex up function of i. Thus if the shape of the the points $(i, \log\{f_i/\binom{t}{i}\})$ for $i = 1, \ldots, t$ is convex up, then the capture probability may vary from one unit to the next. The same is true of the Poisson probability plot; a convex up plot indicates heterogeneity in encounter rates. The two probability plots in Figure 18.1 are convex up, suggesting the presence of heterogeneity in these two datasets. We now consider the construction of alternative estimates that address the issue of heterogeneity.

A lower bound estimate can be derived by considering the probability plots in Figure 18.1. It is constructed by calculating the smallest possible value of f_0 which preserves the convexity of these plots, when the point $(0, \log(f_0))$ is added. The value of f_0 can be determined geometrically, as illustrated in Figure 18.2. This yields $\hat{f}_0 = (t-1)f_1^2/(2tf_2)$ as an estimate for the number of missed units for a dataset with t capture occasions; the lower bound estimate of N is

$$\hat{N} = n + \frac{(t-1)f_1^2}{2tf_2}, \tag{18.6}$$

as first derived in Chao (1987); see also Rivest and Baillargeon (2007) for the derivation of a variance estimate. Letting t tend to infinity gives the lower bound estimate for continuous captures, $\hat{f}_0 = f_1^2/(2f_2)$. In Table 18.1, the columns LB give lower bound estimates $\hat{N} = 216$ and $\hat{N} = 3{,}413$ for the two datasets. They are 30% larger than the M_0 estimates. This highlights the severe bias of the latter when there is heterogeneity in the data. See Hwang and Huggins (2005) and Rivest (2008) for further discussions of this point.

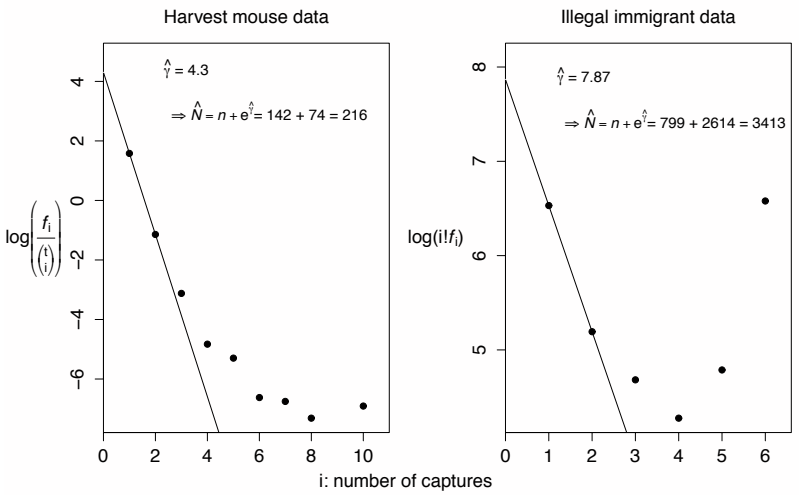

FIGURE 18.2: Geometric construction of the lower bound estimate.

The lower bound estimate can be calculated using formula (18.3) when the parameters (γ, β) are estimated by solving equations that set f_1 and f_2 equal to their predicted values, E_1 and E_2, under M_0 as given in equations (18.1) and (18.5). It makes sense to take out the units caught several times when estimating f_0, especially when the capture probabilities vary between units, because these units might not be representative of the ones that were missed. This increases the variance, however. In Table 18.1, the coefficients of variation (CV) of the two LB estimates are larger than 10%. Observe that in Figure 18.1 the probability plots are nearly linear for $i = 1, 2, 3$; this suggests adding f_3 to f_1 and f_2 to estimate f_0. Bringing f_3 in the estimation amounts to estimating the parameters in (18.1) or (18.5) by fitting the homogeneity model using only the units caught $t_0 = 3$ times or less. This is model M_0 $(t_0 = 3)$ in Table 18.1, where t_0 stands for the maximum number of captures for a unit to be kept when fitting (18.1). This model should give an estimate for N close to the true population size if the units caught three times or less are representative of the units that were missed. The heterogeneity would then be caused by units caught four times or more. They might correspond to a small trap-happy group whose behavior is not representative of the units that were missed.

In Table 18.1, bringing f_3 in the estimation of N reduces both $\hat{N}$ and its standard errors. For the illegal immigrant data the deviance of 6.2 with df $= 1$ indicates that this model does not fit, while for the mouse data the fit is acceptable. The next section discusses estimates of N obtained with models that deals with the heterogeneity by incorporating the random nature of p and λ in the model construction.

18.3.3 Parametric Models with Random Capture Probabilities

In this section, a distribution F_θ describes the variations of the probability of capture p in the population, where θ is a vector of unknown parameters. The choices for F_θ considered here are either the Normal or distributions leading to simple models for the data $\{f_i\}$. As seen in Section 18.3.2, the expectation of the number of units caught i times is

$$\mathrm{E}(f_i) = \binom{t}{i} N E_p\{p^i(1-p)^{(t-i)}\},$$

where the expectation $E_p(\cdot)$ is taken with respect to the distribution F_θ of p. In this context, (18.1) becomes

$$\mathrm{E}(f_i) = E_i = \binom{t}{i} e^{\gamma+\phi_\theta(i)}, \quad i \in \{1,\ldots,t\}, \tag{18.7}$$

where γ is, as before, the logarithm for the predicted number of missed animals and $\phi_\theta(i)$ is a convex function of i that depends on F_θ. The parameters (γ, θ) are estimated by minimizing the deviance (18.2), as in Section 18.3.1. Equations (18.3) and (18.4) are used to estimate N and its sampling variance. For continuous captures, a heterogeneity in encounter rate is modeled by assuming that λ is a random variable with a distribution F_θ and the same construction applies.

The Normal distribution, $\mathcal{N}(\beta, \sigma^2)$, where $\theta = (\beta, \sigma^2)$ are unknown parameters, is often used to model the random capture probabilities and the random encounter rates. This is considered by Coull and Agresti (1999) for the discrete case and by Bunge and Barger (2008) for continuous captures. In Table 18.2, the estimates $\hat{N}$ obtained with Normal random effects are 284 and 7,224 for the harvest mouse and the illegal immigrant examples. The latter estimate is quite large, as it is equal to 2 times the LB estimate of Table 18.1; its high CV of 23% is also noteworthy. Both models fit well as their deviances G^2 are smaller than the degrees of freedom (df). Thus the heterogeneity in capture probability could be described with a Normal distribution.

A simple model can be derived by assuming that, in (18.7), $\phi_\theta(i)$ has a simple linear form such as $i\beta + \psi(i)\tau$ where $\theta = (\beta, \tau)$ are unknown parameters and ψ is a known convex function of i. Lindsay (1986) and Rivest and Baillargeon (2007) show that there exists a distribution F_θ such that $\phi_\theta(i)$ calculated with F_θ gives $i\beta + \psi(i)\tau$ for several convex functions ψ such as $\psi(i) = 2^i - 1$, $\psi(i) = i^2/2$, and $\psi(i) = -\log(3.5 + i) + \log(3.5)$. The corresponding models are called the Poisson2, the Darroch and the Gamma3.5, respectively. Typically one has $\hat{N}(\text{Gamma3.5}) > \hat{N}(\text{Darroch}) > \hat{N}(\text{Poisson2})$, and these estimates allow us to investigate the impact of various specifications of F_θ on the magnitude of $\hat{N}$. For continuous captures, Bunge and Barger (2008) found that Poisson mixtures constructed with a finite mixture of the

TABLE 18.2: Heterogeneity models fitted to the harvest mouse and to the illegal immigrant data.

	Harvest Mouse Data			
	Normal	Poisson2	Darroch	Gamma3.5
$\hat{N}$	284	158	204	350
SE	44	5.1	15	63.9
CV	15.5%	3.2%	7.4%	18.2%
G^2	7.5	77.4	13.9	5.1
df	11	11	11	11
	Illegal Immigrant Data			
	Normal	Poisson2	Darroch	Gamma3.5
$\hat{N}$	7,224	3,374	5,629	9,425
SE	1,651	305	956	2,628
CV	22.9%	9%	17%	27.9%
G^2	.9	3.8	.7	1.0
df	3	3	3	3

negative exponential distribution often provide a very good fit to the aggregated data $\{f_i\}$.

Table 18.2 presents the estimates $\hat{N}$ obtained with these simple models and Figure 18.3 gives the probability plots introduced in Section 18.3.2 and the fitted convex functions, $\hat{\gamma} + \phi_{\hat{\theta}}(i)$, for the four models of Table 18.2. All models appear to fit the data well, except possibly the Poisson2 model for the harvest mouse data. For the harvest mouse data, the fit of the Normal and the Gamma3.5 model are similar, both in Figure 18.3 and considering the deviances of Table 18.2. There is a 20% difference in the two values of $\hat{N}$, however. For the immigrant data, the four models presented in Table 18.2 fit well since their deviances are small. Threefold variations in the value of $\hat{N}$ are observed however and it seems difficult to obtain a definitive estimate for the population size. Figure 18.3 and Table 18.2 illustrate findings of Huggins (2001) and Link (2003): the data at hand does not permit us to identify the distribution F_θ for the random capture probability p (or the random detection rate λ) and the value of $\hat{N}$ depends critically on F_θ. By using auxiliary variables to account for the variations in p and λ, one expects to narrow down the range of possible values for N. This is discussed briefly in the next section.

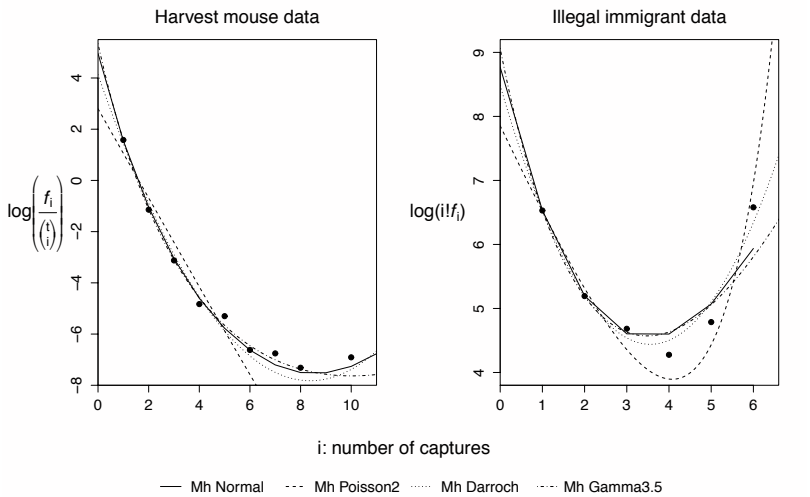

FIGURE 18.3: Graphical representation of the fit of various models for M_h.

18.4 Modeling Capture Probabilities with Unit Level Covariates

This section uses explanatory variables x, measured on each unit caught in the study, to model the capture probability p or the encounter rate λ. Using subscript j for the unit, the dataset is $\{(c_j, x_j) : j = 1, \ldots, n\}$ where x_j is a $d \times 1$ vector. The logit and the log link functions are used to express the capture probability p_j and the encounter rate λ_j in terms of the covariate x_j, viz.

$$p_j = \frac{e^{x_j^\top \beta}}{1 + e^{x_j^\top \beta}} \quad \text{and} \quad \lambda_j = e^{x_j^\top \beta},$$

where β is a $d \times 1$ vector of unknown regression parameters. If p_j and λ_j were known, then N would be estimated using a so-called Horvitz–Thompson estimator, viz.

$$\hat{N} = \sum_{j=1}^{n} \frac{1}{\pi_j},$$

where π_j is the probability of being caught at least once, $\pi_j = 1 - (1 - p_j)^t$ for discrete captures and $\pi_j = 1 - e^{-\lambda_j}$ for continuous captures. By weighting each unit captured by its inverse probability of being caught, one constructs a good estimate of N.

In practice, β is unknown and numerical values for this parameter have to be determined. As in Section 18.3.1, we want to calculate maximum likelihood estimates. This is done by maximizing what is known as the likelihood function. It is constructed by considering the distribution of the number of captures c_j. For discrete captures, the number of captures has a binomial distribution as discussed in Section 18.3.1. Since only units captured at least once are recorded in the dataset, we say that c_j has a binomial distribution truncated at 0. In a similar way, for continuous captures, c_j has a Poisson distribution truncated at 0.

The maximization of the likelihood for discrete captures is carried out with statistical software for truncated binomial regression while one would use a truncated Poisson regression for continuous captures. See Huggins (1989) and van der Heijden et al. (2003) for a detailed presentation of the calculations underlying the estimation of N and of its variance in these two cases. Note, however, that for these complex models the maximization of the likelihood does not produce a statistic, such as the deviance in Section 18.3.1, that could be used to investigate the fit of these models.

If, once the covariates are accounted for, there is still some residual heterogeneity in the data, then the estimate of N might be negatively biased, as in Section 18.3.1. A sensitivity analysis can be carried out by using only the units captured less than t_0 times to estimate the parameter β. As argued in Section 18.3.2, such an estimate should be more representative of the units that were never captured. If the estimate of N is the same as that obtained with the whole dataset, then residual heterogeneity should not be a concern. Such an analysis has been proposed by Böhning and van der Heijden (2009) for continuous captures. To our knowledge, this has not been considered for discrete captures; only models using all the captured units are used to estimate β, regardless of the number of captures c_j.

Table 18.3 gives estimates of N calculated using unit level covariates. Preliminary statistical analyses, not reported here, found that the covariates had a significant effect on p and λ. The capture probability of a mouse increases with its body weight and the encounter rate varies according to the immigrant's country of origin. For both datasets, the pattern of increasing $\hat{N}$ with decreasing t_0 suggests that the covariates do not explain all the heterogeneity in capture probabilities. Since they use all the data to estimate β, the $t_0 = 14$ and $t_0 = 6$ estimates might underestimate the true population sizes in the two examples. Unit level covariates narrow down the range of possible values for N as compared with Table 18.2; note, however, that bringing in the covariates does not seem to reduce the standard error of $\hat{N}$ for estimates having the same magnitude.

The best size estimates of N, among all those considered here, are the ones of Table 18.3 with small values of t_0. However, their CVs are relatively high, especially for the illegal immigrant data, and the lower bound estimate of Table 18.1 might be a better estimate for this dataset. It would be interesting to analyze the data with models featuring both unit level covariates and a

TABLE 18.3: Models with unit-level covariates fitted to the mouse and to the illegal immigrant data.

	Harvest Mouse Data			Illegal Immigrant Data		
	$t_0 = 14$	$t_0 = 4$	$t_0 = 2$	$t_0 = 6$	$t_0 = 3$	$t_0 = 2$
$\hat{N}$	176	197	249	4,883	5,166	5,545
SE	9.4	14.5	40.4	1,114	1,140	1,267
CV	5.4	7.3	16.2	22.8	22.1	22.9

random unit effect. For continuous captures, assuming a gamma distribution for the residual unit effect leads to the zero truncated negative binomial as a distribution for the number of captures C; see Cruyff and van der Heijden (2008). When t is finite, the specification of models for C featuring both fixed covariates and a residual unit effect is more complex and has not, to our knowledge, been investigated.

18.5 Discussion

Capture-recapture methods for closed populations answer a very basic question: How many? This work has treated two examples involving an animal population and an elusive human population, as developments of the statistical methods for this type of data have been motivated by applications in these two areas. Nowadays capture-recapture techniques are applied in a variety of fields. They are, for instance, used in genetic research to estimate the size of a gene pool and in computer engineering to estimate the number of bugs in a new software. Each new application must address the question of heterogeneity: are some units more likely to be captured than others? A failure to deal with this problem may result in a severe underestimation of the population size. This work has presented some of the statistical tools that have been developed to cope with this difficulty and has applied them on two examples.

About the Authors

Louis-Paul Rivest is a professor and the holder of a Canada Research Chair in Statistical Sampling and Data Analysis at Université Laval. He studied mathematics and statistics at the Université de Montréal and at McGill University. His research interests include capture-recapture models, survey sam-

pling, and geometrical and multivariate statistics. He founded Laval's *Service de consultation statistique* and has been involved throughout his career in collaborative research with biologists, foresters, engineers, and social scientists. He was president of the Statistical Society of Canada in 2000–01 and received the SSC Gold Medal for research in 2010.

Sophie Baillargeon is a research associate and lecturer in statistics at Université Laval, where she completed an MSc degree in 2005. She has produced several R software packages that implement methodology developed by the Laval Statistics Group.

Bibliography

Baillargeon, S. and Rivest, L.-P. (2007). The Rcapture package: Loglinear models for capture-recapture in R. *Journal of Statistical Software*, 19. `http://www.jstatsoft.org/v19/i05`, Rcapture CRAN URL: `http://CRAN.R-project.org/package=Rcapture`.

Böhning, D. and van der Heijden, P. G. M. (2009). A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. *The Annals of Applied Statistics*, 3:595–610.

Bunge, J. and Barger, K. (2008). Parametric models for the number of classes. *Biometrical Journal*, 50:971–982.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43:783–791.

Coull, B. A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics*, 55:294–301.

Cruyff, M. J. L. F. and van der Heijden, P. G. M. (2008). Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal*, 50:1035–1050.

Harrison, A. (2012). Counting the unknown victims of political violence: The work of the human rights data analysis group. *Human Rights and Information Communications Technologies: Trends and Consequences of Use*.

Hoaglin, D. C. (1980). A Poissonness plot. *The American Statistician*, 34:146–149.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76:133–140.

Huggins, R. M. (2001). A note on the difficulties associated with the analysis of capture-recapture experiments with heterogeneoous capture probabilities. *Statistics & Probability Letters*, 54:147–152.

Hwang, W.-H. and Huggins, R. M. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, 92:229–233.

Lecren, E. D. (1965). A note on the history of mark-recapture population estimates. *Journal of Animal Ecology*, 34:453–454.

Lindsay, B. G. (1986). Exponential family mixture models (with least-squares estimators). *The Annals of Statistics*, 14:124–137.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59:1123–1130.

McKendrick, A. G. (1926). Application of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). *Statistical Inference from Capture Data on Closed Animal Populations*, volume 62 of *Wildlife Monographs*. Wildlife Society.

Rivest, L.-P. (2008). Why a time effect often has a limited impact on capture-recapture estimates in closed populations. *The Canadian Journal of Statistics*, 36:75–84.

Rivest, L.-P. and Baillargeon, S. (2007). Applications and extensions of Chao's moment estimator for the size of a closed population. *Biometrics*, 63:999–1006.

Rivest, L.-P., Couturier, S., and Crépeau, H. (1998). Statistical methods for estimating caribou abundance using postcalving aggregations detected by radio telemetry. *Biometrics*, 54:865–876.

Rivest, L.-P. and Lévesque, T. (2001). Improved log-linear model estimators of abundance in capture-recapture experiments. *The Canadian Journal of Statistics*, 29:555–572.

Stoklosa, J. (2012). *PL.popN: Population Size Estimation*. R Package Version 1.2. `http://CRAN.R-project.org/package=PL.popN`.

Stoklosa, J. and Huggins, R. M. (2012). A robust P-spline approach to closed population capture-recapture models with time dependence and heterogeneity. *Computational Statistics & Data Analysis*, 56:408–417.

Stoklosa, J., Hwang, W.-H., Wu, S.-H., and Huggins, R. M. (2011). Heterogeneous capture-recapture models with covariates: A partial likelihood approach for closed populations. *Biometrics*, 67:1659–1665.

van der Heijden, P. G. M., Bustami, R., Cruyff, M., Engbersen, G., and van Houweligen, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modeling*, 3:305–322.

Williams, B. K., Nichols, J., and Conroy, M. J. (2002). *Analysis and Management of Animal Populations*. Academic Press, San Diego, CA.