

# WILEY



---

## Fitting the Geometric Distribution to Capture Frequency Data

Author(s): H. Charles Romesburg and Kim Marshall

Source: *The Journal of Wildlife Management*, Jan., 1979, Vol. 43, No. 1 (Jan., 1979), pp. 79-84

Published by: Wiley on behalf of the Wildlife Society

Stable URL: <https://www.jstor.org/stable/3800637>

---

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Wiley and Wildlife Society are collaborating with JSTOR to digitize, preserve and extend access to *The Journal of Wildlife Management*

# FITTING THE GEOMETRIC DISTRIBUTION TO CAPTURE FREQUENCY DATA

H. CHARLES ROMESBURG, Department of Forestry and Outdoor Recreation, Utah State University, Logan, UT 84322  
KIM MARSHALL, Department of Forestry and Outdoor Recreation, Utah State University, Logan, UT 84322

**Abstract:** We computer simulated a mark-recapture experiment where the population size,  $N$ , is known and the recapture frequencies follow the geometric distribution. Two methods described by Edwards and Eberhardt (1967)—linear regression and maximum likelihood estimation (MLE)—were used with these hypothetical data to give estimates of  $N$ . The results were (1) regression gives biased estimates while MLE gives unbiased estimates; (2) both methods produce inaccurate confidence intervals for  $N$ .

J. WILDL. MANAGE. 43(1):79-84

Starting with an assumed process of small mammal capture and home range, Eberhardt et al. (1963) derived the geometric distribution as a model of capture frequencies. In this and subsequent papers (Edwards and Eberhardt 1967, Nixon et al. 1967), rabbit (*Sylvilagus floridanus*) and squirrel (*Sciurus* spp.) capture frequencies were found to fit the geometric model rather well and gave estimated population sizes within a reasonable deviation from the known number of animals. A regression method and a maximum likelihood method are both commonly used to fit capture frequency data (Edwards and Eberhardt 1967). In this paper we examine how good these methods are, i.e., do they give unbiased and precise estimates of population size? We did this by computer simulation. First, we drew a random sample of size  $N$  (each random sample is termed a trial) from a geometric capture distribution based on a known population size  $N$ . Second, estimates of population size and, where possible, confidence intervals for  $N$  were computed using the regression and MLE methods. Third, the average of the population size estimates and confidence intervals were computed from 200 independent trials and judged with respect to the known value of  $N$ . We briefly discuss the use of jackknife estimation, a

technique which has potential for reducing estimator bias and providing confidence intervals.

We express appreciation to D. Hinkley, University of Minnesota, R. Miller, Jr., Stanford University, D. Anderson, D. Otis, and J. Bedford, Utah State University, and the referees for their comments.

This paper was prepared as part of the US/IBP Desert Biome Program (National Science Foundation Grant No. BMS74-02671-A04) and ECOSYM: A Classification and Information System for Management and Wildland Framework (U.S. Forest Service Grant No. 26-3546 and 26-3575).

## METHODS OF PARAMETER ESTIMATION

The geometric model is

$$n_x = Npq^x \quad x = 0, 1, 2, \dots \quad (1)$$

where  $n_x \equiv$  number of animals captured  $x$  times,  $N \equiv$  total number of animals in the population,  $p = 1 - q \equiv$  probability an animal is captured zero times in the trapping experiment.

The data from a trapping experiment consist of values for  $n_1, n_2, \dots$ , up to some maximum value  $n_s$ . The value of  $s$  is the number of trapping occasions, and we assume that  $s$  is "large" so that the effect

Table 1. Means of population size estimates  $\hat{N}$  over 200 trials.

Method	Pop. size $N$	Geometric distribution parameter $p$			
		0.2	0.3	0.4	0.5
R1	50	56.7 <sup>a</sup> (1.0) <sup>b</sup>	48.2 (0.9)	45.9 (1.1)	42.7 (1.2)
R2	50	46.2 (0.52)	44.2 (0.67)	43.5 (0.95)	42.0 (1.3)
MLE	50	50.2 (0.56)	49.9 (0.73)	51.1 (1.1)	52.1 (1.5)
R1	100	92.0 (1.1)	85.3 (1.4)	82.9 (1.9)	82.0 (2.4)
R2	100	91.3 (0.75)	88.8 (1.1)	86.1 (1.6)	84.1 (2.9)
MLE	100	99.3 (0.78)	101 (1.1)	100 (1.6)	99.3 (1.9)
R1	200	169 (2.1)	163 (3.1)	161 (4.3)	164 (4.9)
R2	200	185 (1.2)	179 (2.0)	174 (3.0)	173 (3.9)
MLE	200	200 (1.1)	200 (1.5)	201 (2.0)	201 (2.9)
R1	400	326 (5.0)	322 (6.4)	317 (7.5)	328 (9.8)
R2	400	371 (2.2)	361 (3.4)	351 (4.8)	351 (7.3)
MLE	400	399 (1.7)	399 (2.2)	400 (2.8)	404 (4.1)

<sup>a</sup> Mean of the 200  $\hat{N}$  estimates.  
<sup>b</sup> The half-width of the 95% confidence interval for  $N$  computed from the 200  $\hat{N}$  estimates is given in parentheses. For example, the confidence interval for  $N$  using method R1 with  $N = 50$  and  $p = 0.2$  is  $56.7 \pm 1.0$ .

of truncating (1) on the right is negligible. We cannot, of course, observe  $n_o$ , the number of animals not captured.

Edwards and Eberhardt (1967:92) suggest 2 versions of the regression method (termed R1 and R2 here) and a maximum likelihood method (termed MLE) for estimating the parameter  $N$  in (1).

Both regression methods fit the equation

$$\ln(n_x) = \ln(Np) + x \ln(q) \tag{2}$$

which follows by taking logarithms of (1). For method R1, the slope of the fitted regression equation (2) is an estimate of  $\ln(q)$ , and this can be solved for the estimate of  $p$ . The intercept of (2) is an estimate of  $\ln(Np)$ . This can be solved for  $Np$  and, using the estimate of  $p$ , gives an estimate of  $N$ , written  $\hat{N}$ . For method R2,  $\hat{N}$  is written as the sum of the number of animals not captured,  $n_o$ , and the number of animals  $r$  captured 1 or more times. But on average  $n_o = Np$ , and so the estimate of  $Np$  is obtained from the intercept of the fitted regression. We calculate an approximate confidence interval for  $N$  based on method R2. Seber (1973:172)

briefly discusses the procedure: the fitted regression is used to calculate the confidence limits for  $Np$ , and  $r$  is added to both limits to give an approximate confidence interval for  $N$ .

The MLE method follows Seber (1973:171). Let  $\bar{x}$  be the estimated mean number of captures. Then

$$\hat{p} = (r - 1)/(r\bar{x} - 1) \tag{3}$$

is the minimum variance unbiased estimate of  $p$ . The total population size can be estimated by

$$\hat{N} = r/(1 - \hat{p}) \tag{4}$$

The MLE method provides for an approximate confidence interval for  $N$ . The sample size is  $r$ , and when it is large, as it is in these simulations,  $\bar{x}$  and its standard error can be used to compute an approximate large sample confidence interval for the true mean number of captures. These confidence limits can be transformed by using each confidence limit in place of  $\bar{x}$  in (3) and, in turn, using the computed confidence limits for  $p$  in (4) to obtain the confidence limits for  $N$ . It should be noted that  $r$  in (3) and (4) is

not known and must be estimated from the data. This makes the confidence intervals computed by this procedure approximate.

## SIMULATION METHODS

Simulated realizations of the geometric distribution were programmed on a B6700 Burroughs computer, using the system random number generator, for all  $N$ ,  $p$  parameter combinations in (1) assumed by  $N = 50, 100, 200, 400$ , and  $p = 0.2, 0.3, 0.4, 0.5$ . These settings encompass probable values found in practice. For each parameter combination, 200 independent trials were run. Taking the case  $N = 100$  and  $p = 0.5$  as an example, a trial consists of 100 independent realizations of (1). The number of counts  $n_x$ , for  $x = 1, 2, 3, \dots$ , were then used in each of methods R1, R2, and MLE to give estimated population sizes; the width  $w$  of the confidence interval for  $N$  (upper confidence limit minus lower confidence limit) was computed for methods R2 and MLE. This made up 1 trial, and this was repeated 200 times giving 200 trials. Throughout this study each realization of (1) was statistically independent of all others to the extent that this is possible using computer generated random numbers. We then computed the following summary statistics for each parameter setting based on 200 trials: (1) the mean and standard error of the  $\hat{N}$  estimates computed by methods R1, R2, and MLE; (2) the mean and standard error of the confidence interval width  $w$  computed by methods R2 and MLE; (3) the proportion of the 200 trials in which the confidence intervals based on methods R2 and MLE contained the parameter setting  $N$ . The means and standard errors (items 1 and 2 above) were used to compute confidence intervals for our simulated estimates of  $N$  and  $w$  based on 200 trials.

We and others at Utah State University have checked the random number generator we used, and we believe that the random numbers it generates closely follow the uniform distribution and are not serially correlated. Furthermore, for each simulated trial we performed a chi-square goodness-of-fit test ( $\alpha = 0.05$ ) to see whether there was evidence that the fitted geometric distribution differed from its theoretical counterpart. We rejected the hypothesis of "no difference" in nearly 5% of the trials, or what would be expected by chance if the hypothesis were true. Finally, the results of the simulations show the MLE method gives an unbiased and, over 200 trials, a precise estimate of  $N$ . It is intuitively unlikely that this would occur if our simulated trials were not tracking the geometric distribution.

## RESULTS AND DISCUSSION

Estimates of  $N$  by regression methods R1 and R2 are systematically biased (Table 1), and with the exception of method R1 for  $N = 50$ ,  $p = 0.2$ ,  $\hat{N}$  is less than  $N$ . The largest bias occurs with method R1 and is of the order of 20% of true value.

Edwards and Eberhardt (1967:92) implied that methods R1 and R2 gave equivalent unbiased estimates of  $N$ . This is contrary to our findings. For  $N = 200$ , R2 is less biased than R1; for  $N = 100$ , the 2 methods give nearly the same estimates; for  $N = 50$ , R2 is less biased only for  $p = 0.2$ .

On the other hand, estimates of  $N$  computed by the MLE method appear unbiased. There are 16 parameter settings in Table 1, and only for  $N = 50$ ,  $p = 0.5$  does the MLE method confidence interval not contain  $N$  at the 95% confidence level ( $52.1 \pm 1.5$ ). Given that  $\hat{N}$  is in fact unbiased, we would expect 5% of the cases to produce confidence intervals not

Table 2. Means, over 200 trials, of estimated confidence interval width  $\hat{w}$  and the percentage of the trials in which the confidence interval contained the true population size  $N$ .

Method	Pop. size $N$	Geometric distribution parameter $p$			
		0.2	0.3	0.4	0.5
R2	50	7.7 <sup>a</sup> (0.40) <sup>b</sup> ; 55 <sup>c</sup>	18 (1.4); 78	41 (5.0); 89	
MLE	50	7.6 (0.29); 61	15 (1.2); 72	26 (2.1); 76	
R2	100	13 (0.51); 52	30 (1.7); 76	60 (6.3); 87	110 (11); 91
MLE	100	10 (0.27); 60	19 (0.63); 72	30 (1.0); 81	46 (2.1); 82
R2	200	25 (0.87); 46	53 (2.3); 70	94 (5.3); 84	162 (13); 92
MLE	200	14 (0.27); 63	25 (0.53); 68	40 (1.1); 80	60 (1.9); 81
R2	400	47 (1.4); 48	92 (3.5); 71	157 (7.0); 86	278 (24); 92
MLE	400	20 (0.26); 59	35 (0.50); 77	54 (0.91); 81	83 (1.8); 82

<sup>a</sup> Mean of the 200  $\hat{w}$  estimates.  
<sup>b</sup> The half-width of the 95% confidence interval for  $w$  computed from the 200  $\hat{w}$  estimates is given in parentheses. For example, the confidence interval for  $w$  using method R2 with  $N = 50$  and  $p = 0.2$  is  $7.7 \pm 0.40$ .  
<sup>c</sup> Percentage of the 200 trials in which the 95% confidence interval for  $N$  contained  $N$ .

containing  $N$ ; here, 1 of 16 fails to contain  $N$ . Thus it is reasonable to think that the MLE method produces unbiased estimates of  $N$ .

We do not know the cause of bias with methods R1 and R2, but there are several possible sources. First,  $\hat{N}$  is a nonlinear function of the unbiased intercept and slope estimates and, hence,  $\hat{N}$  is not necessarily unbiased. Second, many of the assumptions implicit in linear regression are violated, and it is possible this causes a lack-of-fit bias: the data are discretely distributed count data, a situation somewhat removed from the assumed normality of errors; the counts  $n_x$  in each  $x$  class are not independent, i.e., a capture count removed or added to a given class must necessarily alter the capture frequency in another class; in equation (2) the error variance of  $\ln(n_x)$  increases as  $n_x$  decreases (Edwards and Eberhardt [1967:92] suggested a weighted regression to ameliorate this). Further, estimation of the standard error of  $\hat{N}$  using method R2 suffers from the same incongruities of reality and assumptions, making confidence intervals for  $N$  inexact.

The 95% confidence intervals produced by methods R2 and MLE capture

the true population size  $N$  in less than 95% of the trials regardless of the values of  $N$  and  $p$  (Table 2). Both methods come nearer to 95% confidence as  $p$  is made larger. The error in the performance of the confidence interval does not appear to be a strong function of  $N$ . As Table 2 shows, the mean width  $w$  of the confidence intervals produced by method R2 is, in general, much larger than that calculated by the MLE method. When  $p$  is 0.4 or greater the confidence interval is wide enough so that even though  $\hat{N}$  is biased, the true value of  $N$  is encompassed by at least 84% of the cases. However, for a confidence interval to be useful for decision-making it must be precise and not biased, and neither of these criteria is met with method R2.

The problem with the confidence intervals computed by the MLE method is not bias in  $\hat{N}$  because  $\hat{N}$  is unbiased. The method we have used for computing MLE-based confidence intervals assumes that the true value of the parameter  $r$  in the transformation equations (3) and (4) is known. For each trial we used the estimated value of  $r$ . In turn, trial-to-trial variation in the estimated  $r$  can cause the computed confidence interval

to be inexact. We present this as a possible, but unconfirmed, explanation.

### Jackknife Estimation

Jackknife estimation technique offers the possibility of (1) reducing estimator bias; (2) establishing confidence intervals in situations where confidence intervals are not obtainable from parametric statistical theory. Mosteller and Tukey (1968:133–160) gave a lucid, nonmathematical introduction to the jackknife technique, and Miller (1974*a*) and Arvensen and Salsburg (1975) reviewed theory and application. To our knowledge, Burnham (1972) was the first to use jackknife estimation for bias reduction in a capture-recapture model. The theory for obtaining jackknife estimates for a function of the estimated intercept and slope of a linear regression equation is due to Miller (1974*b*). Hinkley (1977) extended Miller's work to a weighted jackknife which is less biased when the parameter being estimated— $N$  in this analysis—is a nonlinear function of the regression intercept and slope.

We computed a weighted jackknife estimate of  $N$  based on regression method R1. We found the jackknife estimates of  $N$  were not significantly different from those obtained by method R1 reported in Table 1, i.e., in this case at least, jackknifing does not reduce bias. The potential of a jackknife estimator for reducing bias is restricted to bias resulting from the intrinsic mathematical properties of the estimator. If the bias is due to the inappropriate use of regression analysis (called lack-of-fit bias earlier) then the jackknife estimator would not be expected to reduce bias. The confidence intervals obtained from the jackknife procedure are approximately 85% of the width of those obtained by method R2 (Table 2). But because  $\hat{N}$  remains unbiased, the

fact that confidence intervals are possible for method R1 is somewhat academic.

We did not investigate the possible merits of an MLE-based confidence interval for  $N$  computed by the jackknife procedure. In general, finding the best jackknife estimator is a matter of trial and error. There are arbitrary choices connected with grouping the raw data and data transformation to another scale before the mechanics of the jackknife procedure are executed. We lacked the time for what would necessarily be an extensive study, but believe that a jackknife estimator used with the MLE method could produce better confidence intervals.

### CONCLUSIONS

The simulations of the geometric distribution of capture frequency are a basis for judging the accuracy and precision of regression and MLE estimates of population size. Both regression methods are biased. Regression method R2 is, for  $N \geq 200$ , less biased than method R1. The MLE method is unbiased, and for this reason is preferable to either method R1 or R2. Methods R2 and MLE give confidence intervals that fail to match their theoretical expectations.

Finally, a number of regression-based population size estimators are in the literature, e.g., Seber (1973) details DeLury's, Tanaka's, and Marten's methods of regression estimation. For such methods, the degree with which the assumptions made in the course of the mathematical development of these models depart from reality and the effects on the accuracy and precision of the population size estimates are unknowns. The only way to answer this is by examining the robustness of the estimators against a simulated background in which the true parameters are known.

## LITERATURE CITED

- ARVESEN, J. N., AND D. S. SALSBURG. 1975. Approximate tests and confidence intervals using the jackknife. Pages 123–147 in R. M. Elashoff, ed. *Perspectives in Biometrics*. Vol. 1. Academic Press, New York. 200pp.
- BURNHAM, K. P. 1972. Estimation of population size in multiple capture-recapture studies when capture probabilities vary among animals. Ph.D. Thesis. Oregon State Univ., Corvallis. 176pp.
- EBERHARDT, L. L., T. J. PETERLE, AND R. SCHOFIELD. 1963. Problems in a rabbit population study. *Wildl. Monogr.* 10:1–51.
- EDWARDS, W. R., AND L. L. EBERHARDT. 1967. Estimating cottontail abundance from live-trapping data. *J. Wildl. Manage.* 31:87–96.
- HINKLEY, D. V. 1977. Jackknifing in unbalanced situations. *Technometrics* 19:285–292.
- MILLER, R. G. 1974a. The jackknife—a review. *Biometrika* 61:1–15.
- . 1974b. An unbalanced jackknife. *Ann. Math. Stat.* 2:880–891.
- MOSTELLER, F., AND J. W. TUKEY. 1968. Data analysis, including statistics. Pages 80–203 in G. Lindzey and E. Aronson, eds. *Handbook of Social Psychology*. Vol. 2. Addison-Wesley, Reading, Mass. 819pp.
- NIXON, C. M., W. R. EDWARDS, AND L. L. EBERHARDT. 1967. Estimating squirrel abundance from livetrapping data. *J. Wildl. Manage.* 31:96–101.
- SEBER, G. A. F. 1973. The estimation of animal abundance and related parameters. Chas. Griffin Co., London. 506pp.

*Received 3 November 1976.*

*Accepted 4 August 1978.*