

Statistical Distributions and Population Sampling

The ability to estimate the population parameters of species is an important task in our society. For example, to determine the status of threatened or endangered species, researchers must be able to show their estimated abundance. The same principles used to help track and increase these populations may be used to survey pest populations and determine the best way to destroy them. There are several methods used to analyze populations, with each functioning best in particular a circumstance. The method used is often determined by the density and dispersion of the population, and the type of data that is desired at the completion of the sampling.

The Poisson Distribution

The Poisson distribution is a sampling pattern that is primarily used in populations with random dispersion. This pattern is determined by only one parameter, the mean number of individuals per quadrat. In the Poisson, a change in the mean number of individuals per quadrat directly affects the frequency distribution. At the starting value of four, there is an average of four individuals in each sample unit, and the distribution chart reflects how the actual value for a specific chart may range from zero to eight. After increasing or decreasing the mean, the frequency distribution adapts to display the actual values and how they center around the mean.

The Poisson Distribution: URS

In an unrestricted random sample (URS), a series of samples are taken at random from the population. As stated by Waters, there should be no attempt at grouping or clumping samples, and the result of one sampling should not affect the next. A URS on the *sampling.xls* population gave the results below:

Sum	45	172
	Mean	3.750
	MC	3.822
	IP	1.019

It should first be noted that the *ctl-s* macro did not function, in Excel, even with the Analysis Toolpak – VBA option activated. The program did report: “Descriptive Statistics: Input range contains non-numeric data.” The only consequence was that the recommended random quadrants had to be changed to red by altering the font color of the box.

The results as shown above are a fair estimation of the actual population. The observed mean value is slightly less than the expected mean value. A χ^2 goodness of fit test would probably indicate that the difference was not statistically significant. In other words, it would attribute the deviation to chance. The mean crowding value, as defined below, is approaching 4.

$$MC = m = x = \frac{\sum N_i(N_i-1)}{\sum N_i}$$

Finally, the index of patchiness for the population is very close to one. This indicates that the sample shows a random dispersion in the number of individuals per quadrat. As the actual IP value (1.019) is slightly over one, there may be some aggregation, however this is probably a result of sampling error.

The Poisson Distribution: SRS

The stratified random sample (SRS) is used in cases where the sampling universe has more than one type of habitat, and concurrently, may have unequal distribution of individuals. By dividing the sampling universe evenly in half, two separate habitats of equal size are inferred. Six samples were taken from each habitat, with results very similar to those seen in the URS experiment.

Sum	45	136
	Mean	3.750
	MC	3.022
	IP	0.806

Identical to the URS experiment, the mean value was again 3.75. However, the N_i values used to calculate the mean were different. This resulted in different values for the $N_i * (N_i - 1)$ values, which consequently altered the MC and IP values. The mean crowding value decreased when compared to the URS experiment. The IP value also decreased in the SRS sampling study, and is well below one at a value of 0.806. This suggests that the population is exhibiting uniform dispersion to some extent.

The Poisson Distribution: SS

The final Poisson distribution pattern used to sample is the systematic sample (SS). This type of sampling pattern is simple to design and implement, as quadrats are selected at regular intervals. The results for this sample are shown below.

Sum	51	192
	Mean	4.250
	MC	3.765
	IP	0.886

The SS sample gave results with a higher than expected mean. However this calculated mean value is no closer to or further from the actual mean than in the URS and SRS studies (each was 2.50 units away from the actual). The mean crowding value again approached 4.00, as in the URS experiment. Finally, the IP value was below 1.0, as in the SRS experiment. Thus, again, it suggests uniform dispersion.

Each of the three types of Poisson sampling schemes (URS, SRS, and SS) produced very similar data. This is not surprising in the randomly distributed population simulation, as each of the sample structures are based on a randomly dispersed population. It is likely that the individual differences between these three alternatives will have added significance in alternatively dispersed populations and in natural populations.

Negative Binomial Distribution: URS

A second type of sampling scheme is used specifically for studying aggregate populations. Populations of this type can be seen in species that exhibit gregarious feeding or mating behaviors, for example. Two parameters, the mean and the variable k define the negative binomial. k represents the degree of clumping present in the population. The negative binomial relates to the equation $\sigma^2 > \mu$, where the variance is greater than the mean.

A sample with the original values for the population returned the following data.

Sum	11	22
	Mean	0.917
	MC	2.000
	IP	2.182

The mean is very close to the actual value of 0.95. The MC and IP data points are both positive, and the index of patchiness suggests an aggregate pattern of dispersion.

Changing either of the two terms that define the negative binomial will affect the frequency distribution of the population. For example, the preset mean of 0.95 results in a distribution with a broad range of individuals in each sampling site, from zero to seven. A low mean such as 0.05 will drastically alter this distribution, giving a population of only four individuals with each individual residing in one of the one hundred plots. Altering the k value will affect the clumping in the population. A lowered k -value increases the clumping, whereas an extremely high value can give an IP value that suggests a random or uniform dispersion ($IP =$ or < 1).

Binomial Distribution: URS

The binomial distribution is generally reserved for populations with uniform dispersion. The distribution of the population is determined only by the value of the p term, which indicates the proportion of one outcome. As this type of test is generally used to determine the presence or absence of a species (or population of a certain size), it is more qualitative than quantitative. Altering the p -value thus changes the probability of obtaining one single outcome from 100% ($p=1$), to 27% ($p=.27$), for example.

A sample with a p -value of 0.3 and two possible outcomes returned the following results:

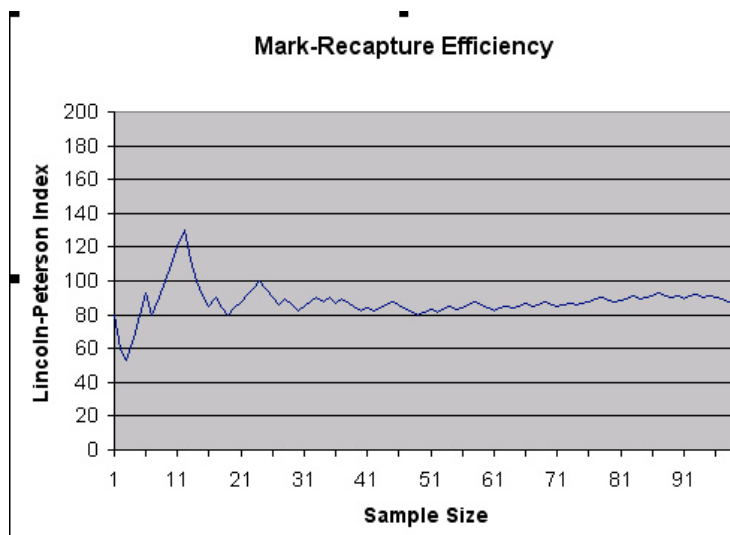
Sum	18	18
	Mean	1.500
	MC	1.000
	IP	0.667

The mean value appears close to an actual value, and the MC value of one suggests uniformity. The IP value also corresponds to a population with a uniform dispersion.

Mark-Recapture

Mark-recapture studies are a type of sampling method that is useful for studying mobile species. The benefit of these types of studies, as reasoned by Krebs, 2001 is that after capturing, marking, and releasing animals, the proportion marked in the following samples should be representatives of the proportion marked in the entire population. The basic approach to this method is the Petersen-Lincoln method, which includes several assumptions. These include obvious assumptions such as a marking procedure that causes no mortality or behavioral affects, and a closed population without dispersal. However, they also include more insightful and specific assumptions such as equal catchability, which are important to consider when forming research projects.

A useful way to represent the results of mark-recapture studies is to graph the Petersen-Lincoln index as a function of the number of individuals in the second sample (N). At the lowest N values, the Peterson estimate is generally low, with a large deviation from 100. At the higher N values, however, the probability of a recapture (R) is also higher, resulting in a more accurate Peterson estimate. Changing the number of individuals marked in the first sample (M) can also greatly affect this graph. As the M value approaches the N value, the graph becomes more linear, as the proportion of marked individuals are recaptured. A lower M value will cause the graph to have more dramatic peaks and valleys, as the number of recaptures fluctuates more intensely.



A Monte Carlo simulation is a convenient way to sample a simulated population. Statistical analyses of the data that results can also help to determine the efficiency of the Petersen-Lincoln method of mark-recapture sampling. The table below was derived from three separate Monte Carlo simulations, (each with P[remain] and P[recapture] values equal to 1).

M	Minimum	Maximum	Mean ± Std Err	Median	Mode
20	41.3	310	98.94 +/- 1.2	88.57	88.57
40	56.36	248	99.43 +/- 0.71	95.38	103.3
60	68.89	169.1	98.48 +/- 0.45	97.89	97.89

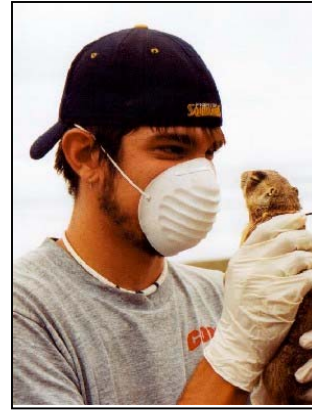
The most obvious trend evident from this data is that the larger the number of individuals marked in the first sample, (M), the more accurate the final data will be. With a low M value, the minimum and maximum values fluctuated greatly, and the standard error was large. The median (central value in a series of data, with an equal number of values above and below it) was fairly low, as was the mode (the most common data point in the series). An increase to 40 pre-marked individuals gave better results, but the most accurate data came from the sampling with an M value of 60. The standard error is reduced, and the median and mode are the closest to the actual value out of any of the samples. This suggests that the Petersen-Lincoln index can be effective and efficient, if the assumptions are held true.

Natural populations are subject to a great deal of diversity due to the wide variability of environmental conditions. Thus the several assumptions used in the basic mark-recapture method are often impossible to maintain. However, even this trend of *failing* assumptions can be replicated to a point in simulations. In the data below, the P(Remain) values represent the percent of the captured individuals that will not leave the study site before being recaptured. The P(Capture) values indicate the percent likelihood that a previously captured individual will be captured again as compared to an uncaptured individual.

M=40	Minimum	Maximum	Mean ± Std Err	Median	Mode
P(Remain) = 0.6	53.91	248	99.82 +/- 0.74	95.38	103.33
P(Capture) = 0.4	44	248	99.22 +/- 0.76	95.38	103.33
P(Capture) = 1.4	53.91	248	99.99 +/- 0.75	95.38	95.38

In the first situation, 40 of 100 individuals were captured initially. However, the P(Remain) value indicates that 40% of the population will leave the study site before they can be recaptured. It seems like this open population situation would have a large effect on the reliability of the data, but it still seems fairly consistent with the expected outcome. In the second example, the P(Remain) value was returned to 1, but the P(Capture) value was set to cause previously captured individuals to become trap shy, and be only 40% as likely to be captured as unmarked individuals. The results from this situation were nearly identical to the previous situation, and indeed, both were affected by a reduced catch rate in 40% of the individuals. In the final example, the P(Capture) value was set at 1.4. In this case, the previously captured individuals are 40% *more* likely to be captured than unmarked individuals. Again, the alterations from the assumptions of the basic model result in data points that are less accurate than seen in the earlier table.

This “less accurate” data obtained from altering the basic assumptions of the mark-recapture model are much more realistic when applied to the natural world. A summer with the USFWS allowed me to see the futility of applying those assumptions to natural populations, prairie dogs in my case. Our telemetry data showed that *Cynomys leucurus* had a wide dispersion range, and our trapping experience showed that some individuals were VERY likely to return to a trap after being caught and tagged previously.



The table below depicts an attempt to estimate a population size of 245 individuals using pre-determined M and C values.

Number of Trials	Maximum	Minimum	Mean \pm Std Err
10	560	93	243.3 +/- 180

The results of this attempt were satisfactory in some regards. With M and C values of 35 and 15, respectively, a mean very close to the actual population size was obtained. However, equally important are the data points that were used to arrive at this number. With extreme maximum and minimum values and a huge standard error, the accurate mean value is reduced in its significance.

Estimating population parameters and understanding statistical distributions are two skills that are widely used throughout the sciences. While every researcher may not use each type of situation, it is important to understand that the variability of nature can affect the results of these studies. Additionally, some use of these parameters and distributions allow a chance to think critically about real world applications, and how solutions to the problems discussed can be achieved (or circumvented).