Examples, Illustrating the Design and Analysis of Monitoring Surveys in National Parks

Paul Geissler
USGS Biological Resources
Paul_Geissler@usgs.gov

Before discussing the design of monitoring surveys, I will consider some background information that will influence our design decisions. Then I will offer some suggestions for the design, illustrated by a simple example. The analysis will also be illustrated using that example. These suggestions are based on my interpretation of the discussions at a workshop (Fancy 2000) organized by Steven Fancy (National Park Service) on February 23-24, 2000 to develop some recommendations for designing a sampling program. The panel members were Paul Geissler, Douglas Johnson, and John Sauer (U. S. Geological Survey); Lyman McDonald and Trent McDonald (West, Inc.); and Anthony Olsen (US Environmental Protection Agency).

Sampling on a Gradient

There are many gradients and environmental differences that influence the distribution and abundance of plants and animals, including elevation and moisture gradients and differences in soil type and prior land use. What animals and plants you find depends to a great extent on where you put your plot or transect. I will illustrative the effect of a gradient on a simple random sample, a compact cluster sample and a systematic sample, using a simple example. The population consists of the numbers 1 through 9, and we want a sample of 3 numbers.

Population: 1 2 3 4 5 6 7 8 9 **True Mean** = 5

Simple Random Sampling (SRS)
There are 84 Possible Samples:
{1,2,3} $\bar{y}$=2.00, v($\bar{y}$)=0.222
{1,2,4} $\bar{y}$=2.33, v($\bar{y}$)=0.519
…
{7,8,9} $\bar{y}$=8.00, v($\bar{y}$)=0.222

$$\bar{y} = \sum_{i \in S} y_i / n$$

$$v(y) = \frac{\sum_{i \in S} (y_i - \bar{y})^2}{(n-1)}$$

$$v(\bar{y}) = \frac{v(y)}{n}\left(1 - \frac{n}{N}\right) = \frac{\sum_{i \in S} (y_i - \bar{y})^2}{n(n-1)}\left(1 - \frac{n}{N}\right)$$

Here y_i is the number in the sample, $\bar{y}$ is the mean, $v(\bar{y})$ is the variance of the mean, n=3 is the sample size, N=9 is the population size, (1-n/N) is the finite population correction factor and $i \in S$ indicates that point i is in the sample S. The expected value of the estimated mean is the mean of estimates from the 84 possible samples: mean (2.00, 2.33, …, 8.00) = 5.00. This equals the true mean of the population, so the <u>estimate of the mean is unbiased</u>. The true variance of estimated mean is the variance of the estimated means from the 84 possible samples around the population mean (5.00): $[(2.00-5.00)^2+(2.33-5.00)^2+\ldots+(8.00-5.00)^2]/84 = 1.67$. The expected value of the estimated variance is the mean of the variance estimates from 84 possible samples: mean (0.222, 0.519, …, 0.222) = 1.67. The <u>variance estimate is unbiased</u>, because the expected value equals the true value. The <u>intraclass correlation is 0.00</u>.

Compact Cluster Sampling

Population: 1 2 3 4 5 6 7 8 9 True Mean=5

Cluster samples are used to reduce travel times between points. Often sample points are located along a transect or subplots selected near a randomly selected point. For our example, there are 3 possible samples:

{1,2,3} $\bar{y}$ =2.00, $v(\bar{y})$=0.222

{4,5,6} $\bar{y}$=5.00, $v(\bar{y})$=0.222

{7,8,9} $\bar{y}$=8.00, $v(\bar{y})$=0.222

The expected value of the estimated mean is mean (2.00, 5.00, 8.00) = 5.00. This equals the true mean of the population, so the <u>estimate of the mean is unbiased</u>. The true variance of the estimated means is $[(2.00-5.00)^2+(5.00-5.00)^2+(8.00-5.00)^2]/3 = 6.00$. The expected value of estimated variance is mean (0.222, 0.222, 0.222)= 0.222. Thus the <u>actual variance (6.00) is larger than SRS variance (1.67), but the estimated variance is a biased underestimate (0.222)</u>. The <u>intraclass correlation (0.85) is positive</u> (Lohr 1999: 138-143).

Systematic Cluster Sampling

Population: 1 2 3 4 5 6 7 8 9 True Mean=5

There are 3 possible samples:

{1,4,7} $\bar{y}$=4.00, $v(\bar{y})$=2.00

{2,5,8} $\bar{y}$=5.00, $v(\bar{y})$=2.00

{3,6,9} $\bar{y}$=6.00, $v(\bar{y})$=2.00

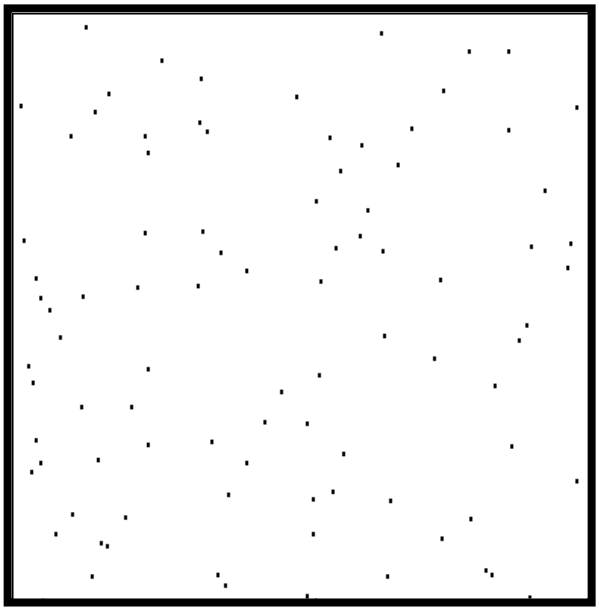
The expected value of the estimated mean of y is mean (4.00, 5.00, 6.00) = 5.00. This equals the true mean of the population, so the estimate of the mean is unbiased. The true variance of the estimated means $[(4.00-5.00)^2+(5.00-5.00)^2+(6.00-5.00)^2]/3 = 0.67$. The expected value of estimated variance is mean (2.00, 2.00, 2.00) = 2.00. Thus the actual variance (0.67) is smaller than SRS variance (1.67), but the estimated variance is a biased overestimate (2.00). The intraclass correlation (-0.35) is negative.

Conclusions

These results are summarized in the following table.

	Simple Random Sample (SRS)	Compact Cluster Sample	Systematic Cluster Sample
Mean estimate	unbiased	unbiased	unbiased
Variance estimate	unbiased	biased too small	biased too large
Actual variance		larger than SRS	smaller than SRS
Correlation	0	positive	negative
Advantages	frequently used	saves travel time	variance smaller than SRS
Disadvantages	inefficient	- variance greater than SRS - variance is biased unless cluster considered	- variance is a conservative overestimate

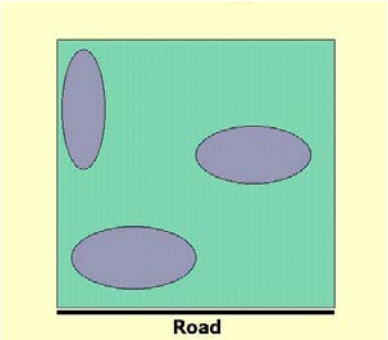
Simple random sampling provides unbiased estimates of the mean and variance. A park can be stratified into more homogeneous areas to assure an adequate sample size in rarer habitats and to increase the precision of the estimates. Within a stratum, simple random sampling, compact cluster sampling, or systematic cluster sampling can be used. Often it is advantageous to use both compact cluster sampling and systematic cluster sampling within a stratum. Sample points are selected systematically with a random start to reduce the variance relative to SRS and to spread the sample points out evenly over the park. The pattern to the right shows the locations of 100 (uniformly distributed) random points. Note that the points tend to clump and leave gaps. A cluster sample (a transect or subplots) can be taken at each of these systematically selected points to reduce the travel time between points. For example, if it



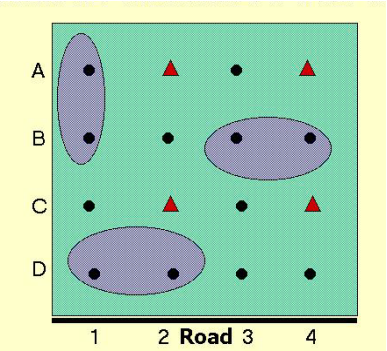
takes three days to get to a point, it does not make sense to only spend 15 minutes collecting data once you get there. However, the data should be summarized (e.g. take the mean) for each cluster, and the variance should be calculated among clusters to avoid underestimating the variance. If a regression or other analysis needs to use the observations from each point of a cluster (e.g., to relate bird counts at each point to the vegetation along the transect), the variance can be calculated using the jackknife procedure (Lohr 1999: 347-368). Using a standard statistical package without modifications will give the WRONG answers.

Survey Design

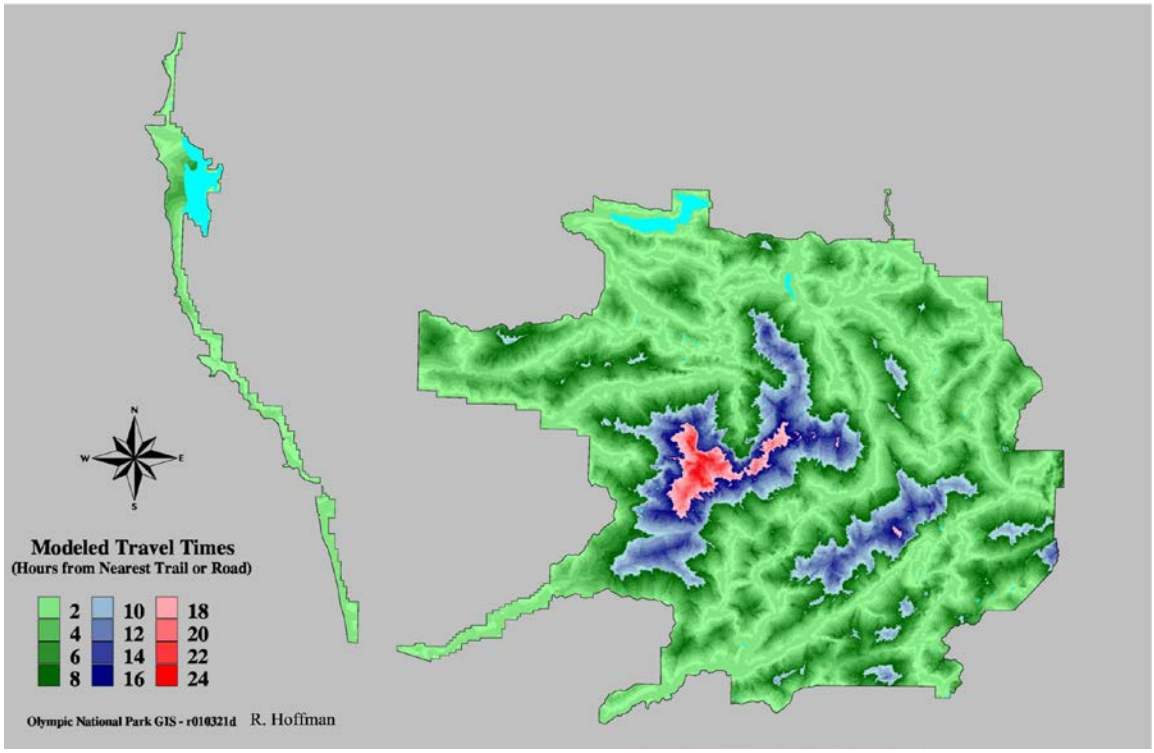
For a simple example, consider two habitat types (green and blue).

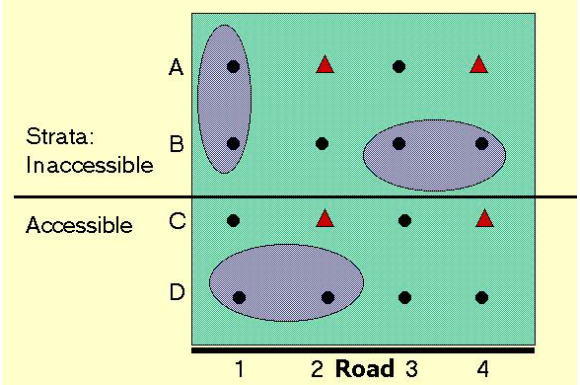


Define a dense base grid (dots) that covers the entire park. Select an initial systematic sample with a random start (triangles) with a sampling intensity that is appropriate for common habitats in inaccessible areas (the minimum sampling intensity). A systematic sample is recommended because it is more precise than a simple random sample. However, both unstratified systematic and simple random samples frequently miss or under sample rarer habitats (blue). Riparian areas are especially difficult to sample because they occupy a very small part of the area of the park. In addition, a systematic (or simple random) sample does not consider the differing costs of sampling in accessible and inaccessible areas.



Roger Hoffman developed the following map of Olympic National Park that shows the travel times to areas of the park from the nearest trail or road. Note that it takes three 8-hour days of hiking to get to some areas from the nearest trail. Sample size and precision can be increased by selecting more points in accessible areas, but some points should be selected in inaccessible areas to provide some information on those areas.



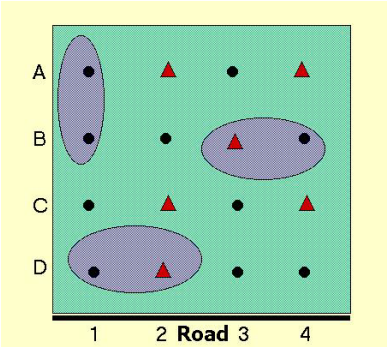


One could use stratification (Lohr 1999: 95-118) to distribute the sample to rarer habitats and to put more sample points in accessible areas. If there is equal interest in all habitat types, then the sample size should be about the same in each to give approximately equally precise estimates for each vegetation type. You may wish to put more sample points in critical habitat types to increase the precision for these habitats. To optimize the sample (minimize the variance of the estimates for the park), considering travel times (costs), the number of sample points in a stratum should be proportional to $N_h S_h \big/ \sqrt{c_h}$ where N_h is the size of the stratum, S_h is the standard deviation and c_h is the cost of sampling (Lohr 1999: 106-113). If information on the standard deviation is not available, and you think it is similar in all strata, make the number of sample points proportional to $N_h \big/ \sqrt{c_h}$. Note that the variance of counts is often proportional to the mean, so that the square root of the expected animal or plant density could be substituted for S_h in the planning, if substantial differences in density among strata is expected.

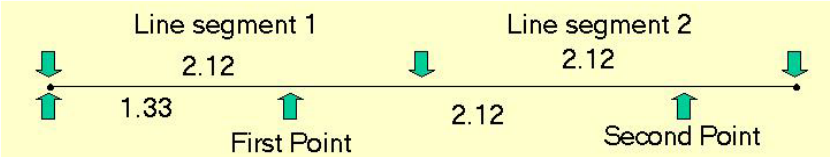
Once drawn, the strata must remain fixed forever. For that reason, it is a good idea to use unchanging features to define the strata and not a vegetation map, which is likely to change. If for example, one defines a stratum to include oak

woodland, but when one arrives at a sample point, one finds an open meadow, the point must NOT be changed to another stratum. A stratum is an area defined on a map for the purpose of distributing the sample, and making any changes will bias the estimation. Although we try to define strata so that they have homogenous vegetation and often name them after vegetation types, strata are logically distinct from the vegetation. Strata are a mechanism to control the selection of the sample with known probabilities and "mistakes" will not bias the estimates, but correcting the "mistakes" will. Domains should be used to make estimates for habitat types, whenever the vegetation does not completely match the strata. I will describe these later.

The unequal probability sampling approach is an alternative to stratification that allows more flexibility and allows changes, although it is more complex. I will illustrate this approach by selecting 2 sample points from the blue areas with probabilities inversely proportion to the square root of the distance from the road (cost).



Point	Dist.	Wt.	Cum.Wt.	Prob.	
A1	4	0.50	0.50	0.12	
B1	3	0.58	1.08	0.14	
B3	3	0.58	1.65	0.14	Selected
B4	3	0.58	2.23	0.14	
D1	1	1.00	3.23	0.24	
D2	1	1.00	4.23	0.24	Selected
Total		4.23		1.00	
Weight per sample 4.23/2				2.12	
Random number a (0<a<1)				0.63	
First point = 2.12 * 0.63				1.33	
Second point = 1.33 + 2.12				3.45	



Think of the weights being laid out on a line 4.23 units long. Divide the line into two equal segments 2.12 units long, one for each sample. Use a table of random numbers to find a random point in the first segment by multiplying the random number (between 0 and 1) by the segment length 0.63(2.12)=1.33. The probability that each point will be selected is proportional to its weight. To find the locations of the other sample points, successively add the segment length (2.12) to the previously selected sample points. This approach uses systematic sampling to increase the precision of the resulting estimates.

If you are following the examples with a hand calculator, note that I did the calculations using a spreadsheet and then rounded the results to simplify the

presentation. Consequently, you will see small rounding errors when you follow the examples with a hand calculator.

For the analysis, we will need the probability of including each sample point in the sample. There were two sampling steps. In the first, we took a systematic sample of 16 possible points. The probability of selecting each was 1/16 = 0.06. In the second step, the probability of selecting each point is given above.

	Prob. Select		Prob. In Sample π
Point	1st step	2nd step	
A2	0.06	0.00	0.23
A4	0.06	0.00	0.23
B3	0.06	0.14	0.42
C2	0.06	0.00	0.23
C4	0.06	0.00	0.23
D2	0.06	0.24	0.55

The probability that a point is in sample = 1 – (probability it was not selected each time). For example,

P(B3 in sample)= $1 - (1- 1/16)^4 (1-0.14)^2 = 0.42$

It is important to sample with replacement to allow the calculation of these probabilities, but with a dense grid there is little chance of picking the same point twice.

Estimation

Using the unequal probability sampling approach, we need the probability that point i is in the sample. As discussed above, it is $\pi_i = 1 - \prod_{c=1}^{C} (1 - p_{ci})^{n_c}$ where there are C sampling steps, and at each step point i has probability p_{ci} of being selected on each of n_c draws with replacement. An estimate of the park mean from a sample point i is $\tilde{y}_i = vy_i / N\pi_i$ where v (≤n) is the number of distinct samples not counting duplicates, y_i is an observation and N is the number of grid points in the park including those which were not selected for the sample (Thompson 1992: 50, 46-53, 67-71). To motivate this transformation, consider simple random sampling without replacement, where π_i = v/N:

$\tilde{y}_i = \frac{vy_i}{N\pi_i} = \frac{vy_i}{N\left(\frac{v}{N}\right)} = y_i$ Here, $\tilde{y}_i$ is just the original observation y_i. Next consider the mean of $\tilde{y}_i$, $\bar{y} = \sum \tilde{y}_i / v = \sum \frac{vy_i}{N\pi_i} / v = \frac{1}{N} \sum \frac{y_i}{\pi_i}$. This is the Horvitz-Thompson estimator of the mean. The estimates of the park mean an its variance are

$\bar{y} = \frac{1}{v} \sum_{i \in S} \tilde{y}$ $\quad s^2 = \frac{1}{v(v-1)} \sum_{i \in S} (\tilde{y}_i - \bar{y})^2 \left[1 - \frac{v}{N}\right]$ and the 100α% confidence interval is $\bar{y} \pm t_{\alpha(v-1)} s$.

For the example, N=16, v=6 and $\tilde{y}_i$ for point A2 is [6(13)]/[16(0.23)] = 21.

Point	Stratum	π_i	y_i	$\tilde{y}_i$	$(\tilde{y}_i - \bar{y})^2$
A2	green	0.23	13	21	82
A4	green	0.23	12	20	115
B3	blue	0.42	55	49	329
C2	green	0.23	18	30	1
C4	green	0.23	17	28	6
D2	blue	0.55	52	35	25
Sum				183	559

Park mean = 183/6 = 31
Variance of mean = [559/(6*5)][(1-6/16] = 12

One can use stratification to increase the precision, making separate estimates for the green and blue areas and then combining these estimates (Lohr 1999: 95-118). Redefine $\tilde{y}_i = v_h y_i / N_h \pi_i$ to estimate the stratum mean instead of the park mean. For example, $\tilde{y}_i$ for point A2 is [(4)(13)]/[(10)(0.23)] = 23. Then a stratum mean and its variance are $\bar{y}_h = \frac{1}{v_h}\sum_{i \in S_h} \tilde{y}_i$ $\quad s_h^2 = \frac{1}{v_h(v_h-1)}\sum_{i \in S_h}(\tilde{y}_i - \bar{y}_h)^2\left[1 - \frac{v_h}{N_h}\right]$

where the subscript h denotes the stratum and $i \in S_h$ indicates summation over the sample units in stratum h. The park mean and variance are

$\bar{y} = \sum_{h=1}^{L} N_h \bar{y}_h \Big/ \sum_{h=1}^{L} N_h \qquad s^2 = \sum_{h=1}^{L} N_h^2 s_h^2 \Big/ \left(\sum_{h=1}^{L} N_h\right)^2$. For the example:

Point	Stratum	π_i	y_i	$\tilde{y}_i$	$(\tilde{y}_i - \bar{y}_h)^2$
A2	green	0.23	13	23	12
A4	green	0.23	12	21	28
C2	green	0.23	18	32	28
C4	green	0.23	17	30	12
Sum				105	80

Stratum mean = 105/4 = 26, variance = [80/(4*3)](1-4/10) = 4

B3	blue	0.42	55	32	34
D2	blue	0.55	52	43	34
Sum				75	68

Stratum mean = 75/2 = 37, variance = [68/(2*1)](1-2/6) = 23
Park mean = [10(26) + 6(37)] / 16 = 31, same as unstratified.
Park variance = [10^2(4) + 6^2(23)] / 16^2 = 5, compared to 12 for the unstratified.

Domains - estimates for a habitat types

Say you want an estimate for the blue vegetation type. This includes points B3 and D2 that are in the blue stratum and point C4 that is in the green stratum, but which was discovered to have blue vegetation when visited on the ground. This is an estimate of a domain or subpopulation (Lohr 1999: 77-81, 60-71). Because the number of points in the domain is a random variable (unknown before the sample was selected), we estimate the domain mean as the ratio of the

estimated park total for the observations to the estimated number of points in the domain, using the transformation to account for the unequal probability of selection. Here S_d refers to the sample points that are in the domain and S refers to all sample points, and v_d is the number of distinct sample points in the domain without counting duplicates.

$$u_i = \begin{Bmatrix} \tilde{y}_i = vy_i / N\pi_i \ if \ i \in S_d \\ 0 \ if \ i \notin S_d \end{Bmatrix}$$

$$t_i = \begin{Bmatrix} \tilde{x}_i = v / N\pi_i \ if \ i \in S_d \\ 0 \ if \ i \notin S_d \end{Bmatrix}$$

$$\bar{y}_d = \sum_{i \in S} u_i \Big/ \sum_{i \in S} t_i = \sum_{i \in S_d} \tilde{y}_i \Big/ \sum_{i \in S_d} \tilde{x}_i$$

$$v(\bar{y}_d) = \frac{\sum_{i \in S_d} (\tilde{y}_i - \bar{y}_d \tilde{x}_i)^2}{v(v-1)\left(\sum_{i \in S_d} \tilde{x}_i \Big/ v \right)^2} \left(1 - \frac{v}{N} \right)$$

$$Confidence\ Interval = \bar{y}_d \pm t_{\alpha,(v_d - 1)} \sqrt{v(\bar{y}_d)}$$

For the example:

Point	Domain	π_i	y_i	u_i	t_i	$(\tilde{y}_i - \bar{y}_d \tilde{x}_i)^2$
A2	No	0.23	13	0	0	
A4	No	0.23	12	0	0	
B3	Yes	0.42	55	49	0.88	317
C2	No	0.23	18	0	0	
C4	Yes	0.23	17	28	1.65	869
D2	Yes	0.55	52	35	0.68	136
Sum				112	3.22	1322
Mean				19	0.54	

The estimated domain mean $\bar{y}_d$ = 112/3.22 = 35. Its variance

$v(\bar{y}) = \frac{1322}{6(5)0.54}\left(1 - \frac{6}{16}\right) = 51$ and the 95% confidence interval is

35 ± 4.303√51 = 35 ± 31. The large confidence interval in this example results from point C4 being very different from the other points.

References

- Fancy, S. 2000. Guidance for the Design of Sampling Schemes for Inventory and Monitoring in National Parks
    - http://science.nature.nps.gov/im/monitor/docs/nps_sg.doc
    - http://science.nature.nps.gov/im/monitor/docs/examples.doc
- Lohr, S.L. 1999. Sampling: Design and Analysis. Duxbury Press.
- Thompson, S.K. 1992. Sampling. Wiley.