

## **Forest & Rangeland Ecosystem Science Center Olympic Field Station**

### ***Statistics of Sampling for Long-term Ecological Monitoring in Olympic National Park:***

#### ***Summary of a Statistics Workshop-April 2-3, 2001***

Andrea Woodward and Kurt Jenkins

USGS Forest and Rangeland Ecosystem Science Center, Olympic Field Station,  
600 E. Park Avenue, Port Angeles, WA 98362.

**Note:** This report summarizes the results of a workshop held at Olympic National Park. These concepts and interpretations have not been peer reviewed and have not been approved by the USGS director. They are provided with the understanding that they are not guaranteed to be complete. **Anyone interested in citing this information should contact the authors before using**, and should carefully consider the provisional nature of this information. Conclusions drawn from, or actions undertaken on the basis of, such information are the sole responsibility of the user.

## **Introduction**

The USGS-BRD team at Olympic National Park is helping the park build a long-term ecological monitoring program. Our goals are to create a conceptual structure for the design of monitoring, to explore sampling variation and determine attributes that can be sampled cost-effectively, and to develop an umbrella sampling plan including an integrative and feasible sample design. Developing an umbrella-sampling plan is an important initial step in developing a monitoring program to ensure spatial integration of individual monitoring components. Further, the evaluation of sampling variability is a critical step in selecting attributes of natural systems to monitor and to avoid investing in costly or fruitless monitoring ventures. Based on these premises, we recently completed three years of data collection to support recommendations we will make to the park. We have data sets describing annual and spatial variation of several animal species, plant communities and populations, as well as data evaluating grid-based versus stratified random sampling schemes.

In recent years we have become aware of changing perspectives of how best to determine the adequacy of a sample to detect change. Early on we planned to use regression-based methods to examine power of various sampling efforts to detect a specified effect. More recently, we have thought about detecting change from a repeated-measures ANOVA perspective. Most recently, we have become aware of paradigms not grounded in a traditional hypothesis-testing framework. For example, we might evaluate sampling sufficiency to detect extreme or unusual values in a data set over time.

Consequently, we decided to convene a small group primarily of biometricians and members of the park's resource management staff to discuss theoretical and practical answers to the following questions:

*What is the question?* How do we frame the question of change detection?

*How much sampling is enough?* When is power analysis useful? What other means are there to determine 'n' *a priori*?

*Where shall we sample?* What is an effective sample frame for distributing plots on the landscape?

*How often is enough?* What temporal sampling frames might we consider given observed values of annual variation?

The workshop was characterized by a lively and productive discussion around all of these questions. Although focused on the particular circumstances at Olympic National Park, many of the answers have more general applicability. The answers to each emerged throughout the meeting so the meeting will be summarized by question rather than chronologically. A list of workshop participants is given in Appendix A.

## **What is the question?**

Paul Geissler and Eric Rexstad addressed this question by each listing methods commonly used to detect change, then giving examples of power analyses based on a

traditional hypothesis-testing framework (Geissler) and methods developed to identify extreme values in data time-series (Rexstad).

*Standard Methods Used to Detect Change*

- Compare Period Means – This approach is easily understood but cannot be used if sites or methods change among periods. You judge that there has been a change using years within periods as replicates. That is, if the difference between periods is more than you would expect based on the within period differences, then you conclude there is a change.
- Regression: Linear, Exponential, Multinomial or Route – Linear regression rarely fits the data well, but is used to estimate an average rate of change. It can be used when sites or methods change. Usually a log transformation is used to fit an exponential regression model, which may better detect growth or decline of populations with constant rates of change. To accommodate the repeated annual measurements on the same transect (route), the route regression approach fits a linear regression on the log scale (exponential model) for each route and then uses the fitted slopes (trends) as the observations in a stratified random survey design. With the route regression approach, you judge that there has been a change using the sites as replicates rather than using years as in the period means approach. That is, if you get the same answer (e.g., a positive change) at several sites, then you conclude there is a change. However, if the rate of change is not constant during the time-series a curved line such as a polynomial may be closely fit to the data, but the line doesn't measure change.
- Rank-Based (Non-parametric) Statistics – These methods enable analysis of data that does not meet the assumptions required by parametric statistics. It is always prudent to check whether the data meet assumptions of the statistical model. For example, with the route regression model, the route slope estimates should be checked for normality, constant variance and outliers. While transformation can often normalize observations and stabilize their variance, outliers can substantially change the results, and they should not be deleted unless they are mistakes. Nonparametric methods can reduce the influence of outliers and do not require normality assumptions. However, nonparametric methods are often not as flexible as parametric methods.
- Permutation Methods (Computer intensive) – These methods involve repeated re-sampling of the original data, and have only been feasible since the advent of personal computers. These methods are often useful when it is necessary to use data from individual points within a cluster sample as the observational unit. For example, one may want to estimate bird-habitat relationships using point data along transects, but still calculate the variance among transects, necessary because the points are not independent.

*Analysis of Channel Islands Exotic Plant Monitoring – Paul Geissler*

Paul Geissler used a sample design called a split block, which includes attributes of both a split plot and block designs and is analogous to an agricultural split plot design because treatments are not randomly assigned to each cell (see Appendix B). It includes repeated measures in time and space:

	Period A		Period B	
	Yr 1	Yr2	Yr3	Yr4
Plot A				
Subplot 1	X	X	X	X
Subplot 2	X	X	X	X
Plot B				
Subplot 3	X	X	X	X
Subplot 4	X	X	X	X

The analysis is a paired-t test or ANOVA.

In the specific example of the Channel Islands data, change was evaluated between two time periods, early (pre-1996) and late (1996-2000), and each year was categorized as having low (< 13 in) or high (> 13 in) precipitation. The result was a split block in time with sites as blocks (or reps) and precipitation categories nested within periods. The resulting ANOVA had Site, Year, and Period, Precipitation and Period x Precipitation within Year as sources of variation, and Site x Year as the error term. (See Appendix C for examples.) The objective of the analysis was to test for change through time and illustrate the importance of precipitation.

Years at Channel Islands were divided into early and late periods simply by dividing the data from the site with the fewest years of observation into two equal periods. That break point was then used for all sites even though some had a longer record, and therefore more years in the early category than the late. One could also (more meaningfully?) break time periods by, for example, occurrence of a management action, documented climate shift, or catastrophic event.

The original data were used to calculate least-square means (necessary because of unequal sample size) and power. The data were transformed before calculating the ANOVA.

An important observation about the ANOVA approach used at Channel Islands is that it is a special case of the general linear model (regression, analysis of variance, analysis of covariance, etc.), the standard method for determining relationships among variables. Because the analysis assumed an underlying relationship of plant cover with time period and precipitation, the sample points were assumed to have equal probability of sampling, and the error terms met several necessary assumptions, it was appropriate for “model-based” analysis. Model-based regression is the traditional version taught in beginning statistics and used in statistics packages such as SAS and SPSS. The variance estimate is based on deviations of observations from the model, and inference can be made to universal model parameters that describe more than the sampled population.

An alternative for survey data is the design-based analysis. There has been much debate among survey statisticians on the advantages of design-based and model-based analyses. In the design-based approach, the random process comes from the selection of the sample and the tests come from the randomization, making no assumptions about the distribution of the data. The model-based approach assumes some model, without specifying how points are chosen. The test is how well the model predicts the observations, with the randomization coming from the data. Many statisticians use a combination of these

theories and methods: models are often used to analyze data from design-based surveys, and randomization is often used to select samples for model-based studies. Geissler was especially critical of a pure model-based approach for monitoring that assumes that one can adequately model all the environmental and other factors affecting the distribution and abundance of animals and plants. Of course, if this were true, one would not need randomization. From the design-based perspective, complex survey data, whose sampling structure may include stratification, unequal probabilities, and clustering, do not meet the assumptions needed for model-based regression. Often the probability of selection of a data point is related to the response. For example, the sample frame may target rare plant associations and sample them more heavily than they would be by a purely random sample of vegetation. Consequently, complex survey data should be analyzed using specialized software (e.g., PC CARP, OSIRIS, SUDAAN) because model-based inferences about parameters and variance will most likely be wrong. Standard statistical packages such as SAS and SPSS can be used only if appropriate transformations and procedures are used to “trick” the packages into providing the correct answers. (See Lohr Chapter 11 for more information)

*Identifying Extremes in Ecological Data – Ed Debevec and Eric Rexstad*

The standard methods for change detection are effective for populations that undergo directional changes, but are not effective for eruptive, cyclic, or highly variable populations such as microtines. Yet microtines are thought to be potentially sensitive indicators of environmental change because their small home range, high metabolic rate and rapid population turnover might make them sensitive indicators for changes in climate or trophic dynamics. Consequently, it is important to find a solution to monitoring these challenging populations.

Instead of focusing on changes in the population mean through time, Debevec and Rexstad have developed methods to identify when an observation in a time-series of a monitored variable fails to conform to the previously measured range of natural variation. Debevec and Rexstad built on previous literature (cited below), which identifies and separates two components of variation in time-series data: ‘process’ variation, or  $\sigma$ , is a measure of inherent variability in the process of interest among years, whereas ‘sampling error’ describes measurement error independent of process. Regarding microtines, we can think of their population numbers as a stable “system of chance causes” whose variability we want to quantify so we will know when an observation is out of the stable pattern. In a more general sense, this approach to detecting changes in long-term data series has application whenever there is need to determine whether change measured for an attribute transgresses the normal bounds (or range) of ‘natural’ variation as determined from prior data.

The universe of ecological monitoring includes:

- Measuring a single attribute at a single instant (= status)
- Measuring a single attribute repeatedly over a span of time at one location
- Measuring multiple attributes repeatedly at a location over time
- Measuring a single attribute over time at multiple locations
- Measure multiple attributes over time at multiple locations

Rexstad began by considering a single attribute measured over time in one location. The data include observations  $y_1, y_2, y_3 \dots y_n$  with associated standard errors  $se_1, se_2, se_3 \dots se_n$ . The standard error for each year also has inherent variability over time. So solving iteratively for sigma of the standard errors ( $\sigma_{\text{hat}}$ ) will describe process variability. If the system stays within this sigma of sigmas (or variance of the variance estimate), then the system is considered stable. Otherwise the system has changed and we need a statistic that will compare the next observation with the previously established distribution of observations (having mean  $\mu_{\text{hat}}$ ) based on  $\sigma_{\text{hat}}$  to let us put a probability on that change.

A t statistic could work but it must be corrected for the fact that the variance of observation  $n+1$  includes process variation and measurement error. So the appropriate statistic is:

$$t = \frac{y_{n+1} - \mu_{\text{hat}}}{\sqrt{(\sigma_{\text{hat}}^2 + se_{n+1}^2)}}$$

Then the Probability of Conformity (PC) that the latest observation is from the previously described distribution is:

$$PC = 2P[t_n \geq |t_{n+1}|] \text{ where } t_{n+1} \text{ is one-tailed}$$

See Appendix D for a numeric example.

This method requires at least three years of baseline data before the variance of the variance can be described, and the more baseline years, the more accurate the description. The method also assumes that process variation is normally distributed.

Rexstad presented the results from 100-year simulations illustrating that the metric responded appropriately and dramatically to the following situations:

- Various scenarios
  1. When no change occurs the PC shows no response
  2. Change-point perturbation (e.g., Nile River before and after dam)
  3. Trend – incremental creep, PC detects trend even though it learns from previous behavior
- Test with different levels of process (P) and sampling (S as set by CV) variation (all combinations of  $P = 5, 10, 15$  and  $S = 0.05, 0.10, 0.15$ )
- Changes to  $\mu$  and  $\sigma$ 
  1. Increase/decrease  $\mu/\sigma$
  2. Apply perturbation in various years
  3. Apply multiple perturbations in different years
  4. Vary recovery time
- Autocorrelation

Rexstad also showed that this method works for the other monitoring scenarios described above: multiple attributes measured at multiple places or any other combination. As long as test statistics are independent of one another they can be pooled. Then the PC is distributed Chi-squared with degrees of freedom equal to twice the number of data sets. The metric successfully detected a change in one of four data sets, and changes in two of

four data sets when the changes occurred at different times. The metric is also robust to violation of the assumption of independence, either a positive or negative correlation among data sets.

Several summary comments were made about using the PC metric:

- It is best to use it with annual data.
- It can be used for periodic data but it will take longer to show a response to change
- No false positives have occurred in any tests.
- The metric needs some minimum “warm-up” period: 3-5+ years.
- PC is a cousin of “cusums”. This measure determines that probability that a manufacturing process is operating outside acceptable bounds. It falls within the area of statistics known as quality control.
- Circumstances when it should not be used probably exist but they haven’t been identified yet.
- The metric assumes no knowledge of the system that might suggest when action might be taken – if this knowledge exists, it can help determine management thresholds.

So, what is the question? Well, analytical methods exist that allow you to frame whatever question you like, depending on your objectives and the inherent properties of the attribute! The two classes of analyses described above pertain to two fundamental types of questions that could be asked. The classic scenario of power analysis is appropriate to answer questions about sampling requirements to detect changes of various magnitudes and probabilities of error given an underlying variance structure. This approach may work best with populations where biologically significant change is expected to be feasibly detectable despite annual variation and when measurements are not necessarily made annually. The analysis of probability of conformance applies to questions about whether a measurement deviates from the normal range of variation. Although useful for any time-series, it is especially effective when the normal range of variation is eruptive or cyclical and annual measurements are made. These two types of questions broaden the conceptual basis of selecting and evaluating potentially useful indicators for monitoring to include those with high annual variability.

### **How much sampling is enough?**

Because monitoring is all about detecting change, we want to design our sampling to especially avoid making Type II errors, or missing a change when one has actually occurred. If we know the variance of the attribute we are monitoring and the amount of change considered important to detect (i.e., effect size), we can determine the power (the probability of not making a Type II error) of a test to detect that change. Because variance is related to sample size, and power is inextricably linked to effect size, variance and alpha (the probability of a Type I error or detecting a change when none has occurred), one can examine relationships among power and sample size, specified effect sizes, or alpha.

Using power analysis to determine ‘n’ based on the variance observed during pilot studies has become an important tool for designing sampling strategies for monitoring or

for adjusting sampling effort in long-term studies. Several canned programs and statistical software packages are available to examine power of virtually any statistical test used to detect change over time. The reference by Cohen (1988) is a 'bible' of sorts for those wishing to examine power of many complex analysis of variance designs or two sample comparisons (for example comparing two time periods). The programs TRENDS (Gerrodette 1987) and MONITOR (Gibbs et al. 1998) are commonly used to examine the power to detect linear or exponential trends in time-series data.

Paul Geissler demonstrated the use of power analysis in the context of comparing means of vegetation measurements between two periods of time in the Channel Islands (Appendix C). Geissler presented a 'power curve' that relates power of the test to discriminate a true difference to the magnitude of true difference between period means. The graph in Appendix C demonstrates that the replication of sample plots was sufficient to detect a 10% difference in mean cover of exotic plants with 80% certainty.

Kurt Jenkins presented a data set on bats, which he had analyzed using MONITOR, as a basis for initiating a discussion of linear trend detection. In the example, Jenkins had monitored bat activity levels in 6 old-growth Douglas-fir stands over 3 years in each of 2 watersheds. The analysis of power indicated that about 14 such stands would have to be monitored to detect a 5% per annum change in bat activity over 6 years at the watershed level. A discussion of the analysis ensued. Program MONITOR evaluates whether the mean slope of the trend lines fit to each plot's data (either linear or exponential) differs from zero. The program requires several restrictive assumptions about the data. It does not use the data to their full potential because it does not differentiate among annual, spatial, and measurement error although they mean different things. Also, there is a debate in the literature as to the appropriate measurement of within-plot sampling variation. Program MONITOR requires input on the coefficient of variation of slopes measured in independent study plots, although such data is generally poorly understood and difficult to obtain from short-term studies. If slopes of some plots were positive and others negative, MONITOR would see no overall trend and would indicate that additional samples are needed, when in fact, interpretation of the spatial patterns may lead to a different conclusion.

The positive aspects of power analysis are that it gets you into the ballpark of an adequate sample size and it is useful for evaluating the effectiveness of on-going monitoring. So it is a valuable tool for helping you collect sensible amounts of data and to think about the analysis before you collect the data. It also facilitates mid-course corrections, a process that is absolutely critical to a monitoring program.

If relative sampling requirements (based on sampling variability) are used as a criterion to select attributes of natural systems to monitor, attributes with high process variation may fare poorly in a comparison of power if one is estimating trends (e.g., microtines or other cyclic or stochastically variable data). It is useful to keep in mind that it may still be possible to measure attributes with high process variation, but it may be necessary to frame the monitoring question in terms of recognizing when process variation exceeds normal ranges.

In summary, it is often useful to consider sampling requirements when designing a monitoring program and for periodic evaluation. Power analysis is especially important in

the detections of population declines, based on hypothesis tests. Retrospective monitoring of general ecosystem integrity often involves monitoring suites of environmental variables co-located at monitoring sites. While it may be possible to optimize sampling effort (e.g., number of independent sampling sites) for each variable independently based on the sampling variation of each, it will be very difficult to optimize sampling sufficiency for several variables measured at the same sites. For example, abundant species may require fewer plots to achieve adequate power than rare ones. Finally, remember that anything determined *a priori* is a guess, and that all aspects of the sampling should be periodically reevaluated using your experience in the field. In addition to being important for evaluating how effectively your monitoring is detecting population changes, observed power is an important consideration in allocating monitoring resources.

**Where shall we put monitoring plots?**

We have been wrestling with the issue of choosing a sample frame. Intuitively, most biologists and ecologists gravitate toward a stratified random sample to distribute plots on the landscape and to be able to draw inferences to biologically meaningful categories (e.g., vegetation types). Statisticians often now encourage a systematic sample. Given the terrain and remoteness of some areas of the park, we also have to consider constraints due to accessibility. Finally, some monitoring attributes require more intensive measurements than others, and therefore must be done in fewer and/or more accessible places. Our discussion took all of these issues into consideration.

The descriptions and properties of different types of samples were provided by Paul Geissler. With simple random sampling, each point is randomly selected independently of the other points. A compact cluster sample starts with a simple random sample and at each point a cluster of samples is taken (e.g., subplots or a transect). With systematic sampling, sample points are evenly spaced, often on a grid after a random start. Their properties can be summarized this way (see Appendix E for numerical examples):

Sample Type:	Bias of Estimate of Mean	Estimated Variance <sup>1</sup>	True Variance of Estimated Means <sup>2</sup>
Simple Random	Unbiased	Unbiased	= Simple Random
Compact Cluster	Unbiased	Underestimates	> Simple Random
Systematic Cluster	Unbiased	Overestimates	< Simple Random

<sup>1</sup>Estimated variance is unbiased if the expected variance – (mean of variances from all possible samples of a finite population) equals the true variance

<sup>2</sup>True variance is the variance of estimated means drawn from all possible samples of a finite population

Comparing the different types of sample, one important point is that systematic samples are desirable because they have lower variance than simple random samples. Cluster samples, (e.g., five plots are measured at one site) are attractive because you can collect more data for a given amount of travel. However, the five plots are not independent of one another and their variance must be calculated among sites. Therefore, you must use the cluster means for analysis instead of the individual plots within the cluster.

Consequently, the estimate of a population parameter is less precise than with the same number of independent samples. A systematic sample made up of clusters, rather than individual points, is often advantageous.

Any of these methods of selecting a sample (simple random, compact cluster, or systematic cluster) can be combined with any of the following methods of distributing the sample (equal probability, stratified, or unequal probability). With equal probability sampling, all areas are equally likely to be selected. With stratified sampling, the park is divided into relatively homogeneous areas call strata. Equal probability sampling is used within strata, the selection probability and sample density can be different for different strata. With unequal probability sampling, the probability of selection and sample density can vary continuously across the park. Stratified sampling is a special case of unequal probability sampling.

	Pros	Cons
Equal Probability	<ul style="list-style-type: none"> <li>• Simple to implement</li> <li>• All areas are equally important</li> <li>• Emphasizes common species</li> </ul>	<ul style="list-style-type: none"> <li>• Can be inefficient</li> <li>• Provides little information on less-common species</li> </ul>
Stratified	<ul style="list-style-type: none"> <li>• Sample density can be increased to provide adequate samples for less-common species</li> <li>• Sample density can be increased in more accessible areas to increase sample size</li> </ul>	<ul style="list-style-type: none"> <li>• More complicated than equal probability sampling</li> <li>• Strata must remain fixed forever, although one can switch to unequal probability sampling, which will allow changes</li> </ul>
Unequal Probability	<ul style="list-style-type: none"> <li>• It has the advantages of stratification without need to define discrete strata</li> <li>• One can add samples without regard to the initial strata</li> <li>• Probability of selection can vary continuously</li> </ul>	<ul style="list-style-type: none"> <li>• More complex than stratified sampling</li> <li>• One must keep track of the selection probabilities</li> </ul>

Andrea Woodward presented the results of a pilot study evaluating systematic and stratified random sampling frames in a selected watershed of Olympic National Park. Kurt Jenkins presented a ‘straw-man’ sampling frame for consideration as a means of integrating extensive monitoring projects with intensive monitoring projects while incorporating accessibility strata. After considerable discussion, a consensus began to emerge on key features of a sampling frame that would meet the many considerations working in a large wilderness park with limited access:

- Stratify the park by accessibility and black out those parts of the park that cannot be sampled for whatever reason. The blacked-out areas may vary with project. Kurt Jenkins suggested (and showed maps of) accessibility zones: high = within 1.5 km of a road, moderate = within 1.5 km of a trail, low or inaccessible = everything else and

everything with  $>35^\circ$  slope. The key feature is to explicitly identify the sampling universe for each project and the area for inference. Extrapolation to blacked-out areas can be accomplished using professional judgment with the caveat that while adequate for some purposes, professional judgment will be a weak link in controversial decisions, particularly those involving legal challenges.

- Keep track of the probability of a point occurring in a blacked-out area. If some points should become accessible due to changes in technology or location of boats, for example, you can sample them, keeping them separate from the rest of the sample until you are sure they represent the same thing.
- Overlay the park with a dense grid of points: say, “5000 points of light”. This grid can be sampled differently for different purposes. Measurements that can be done extensively (e.g., aerial photography, presence/absence) might be done over the entire grid. More intensive projects might be done only within the easily accessible zone, or if necessary, on a probabilistic sample of grid points. Effects monitoring for management actions should occur on a targeted sample. Hence blacked-out areas will differ among projects.
- Other useful grid strata in addition to accessibility might be east-west regions (to reflect the precipitation gradient) and elevation bands. Never stratify on the variable you are measuring.
- Samples can be allocated on the grid using strata, and domains can be created later for analysis. For example, vegetation categories may occur across strata. However, points for each vegetation category can be combined into a domain by weighting them according to the probability associated with the stratum they come from. (See Appendix F)
- Samples can be added to grid strata by re-sampling the grid. The sampling must be with replacement and the probability associated with the new points will be determined by the intensity of the second round of sampling.
- Co-locate projects as much as possible to maximize ease of interpretation.
- Integrate terrestrial and aquatic sampling. We discussed using hydrologic unit categories (HUC’s) as used in the Aquatic/Riparian Effectiveness Monitoring component of the Northwest Forest Plan for sampling watersheds. However, it was concluded that although HUC’s form a logical sampling unit for riparian areas, they might not be as useful for terrestrial monitoring. While it is important to draw linkages between aquatic and terrestrial monitoring, it may be more informative to monitor terrestrial components associated with streams in transects away from those streams selected in aquatic monitoring.
- There is no flexibility in sampling – all samples must come from the point where they were intended to be, no matter what the characteristics of that site.

Several other insightful comments were made regarding sampling design that will help us conduct effective monitoring:

- Permanent plots are preferred over temporary ones because they are better able to detect change. However, a plan must exist for rotating plots out of the sample because plots will inevitably wear out. If plots are treated differently than adjacent areas in

any way, the results will be biased. Rotating plots out of the sample in this case will reduce bias, but will also reduce the power to detect long-term changes.

- It is desirable to have both extensive and intensive measurements for each component of monitoring. For example, population indices might be measured extensively while population estimates can only be measured intensively and are often used for model development. Intensive studies are frequently subject to sampling bias because they only represent a small area as they cannot be located away from roads; extensive studies are subject to measurement bias because with indices you don't know exactly what you are measuring. Intensive, extensive and manipulative studies are all essential to a good monitoring program.
- We need to have enough climate data so that we can extrapolate across the entire park.
- Avoid bias by visiting sites randomly in time instead of, for example, always doing the easy sites first then running out of time for the harder ones.
- Consider picking indicators that can be "cross-dated", meaning choose indicators from different disciplines that one predicts to detect the same perturbations to the system. For example, an increase in temperature = reduced numbers of bats = reduced cover of *Astragulus*, etc. This will create integration in the monitoring system.
- Sampling efficiency could be increased by designing sampling that can be done while traveling to the grid sites. This sampling could be included in the sampling design as a separate frame. Although biased, the observations might increase precision.

### **How often is enough?**

Andrea Woodward presented three years of data from permanent vegetation plots to discuss how best to distribute samples through time. Because vegetation has high annual variability primarily due to changes in weather, her approach was to average the results over three years to estimate the average condition for period one. At some time interval later, possibly ten years, she would collect another three years of data to characterize period two. A difference between period one and period two would be detected using a paired-t test of period means by plots within vegetation categories with error estimated by the variation of plots within habitat. Her overall strategy was inspired by Lessica and Steele (1996), but it was analyzed substantially differently.

This approach seemed to meet with approval and many helpful comments were made during the discussion. It was noted that the most powerful way to detect change over a long period of time is to put all of the sampling effort at the end points rather than collecting data in the middle. The number of years collected at each end should be determined relative to underlying dynamics. For example, climate exhibits a bi-annual pattern in precipitation so it would be ideal to collect a multiple of two years of initial data and then again at some time interval later. It might also be productive to use weather as a covariate, as demonstrated in the Channel Islands analysis. Additionally, this method will detect a change even if there is high variability in abundance of a taxon among sites, assuming that taxa behave similarly regardless of initial cover. However, there is no information on population fluctuations during the middle of the period.

Periodic sampling may be effective for organisms such as plants that are expected to show directional changes. It may not be effective for organisms such as microtines,

which have eruptive, cyclical population dynamics. In this case, annual data and the technique for detecting extreme values described by Eric Rexstad might be more appropriate.

We did not address the question of how to sample the entire frame over time. For example, how might we allocate our annual effort across plots given that they are grouped by stratum, must be sampled for three consecutive years, and cannot be done all at once? Also, how do we rotate plots out of the sample so they do not “wear out”? These questions will have to wait for another time.

### **References** provided by Paul Geissler and Eric Rexstad

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Assoc. Chapter 8, case 2 describes the use of tables in the chapter for calculating power for main effects in complex analyses of variance.

Gerroddette, T. 1987. A power analysis for detecting trends. Ecology 68: 1364-1372.

Gibbs, J.P., S. Droege, and P. Eagle. 1998. Monitoring populations of plants and animals. BioScience 48:935-940.

Gilbert, R.O. 1987. Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold, New York.

Lessica, P. and B. M. Steele. 1996. A method for monitoring long-term population trends: an example using rare arctic-alpine plants. Ecological Applications 6:879-887.

Lohr, S. L. 1999. Sampling: Design and Analysis. Duxbury Press. Excellent and understandable sampling text which covers both design- and model-based approaches, unequal probability sampling and variance estimation for ANOVA/regression and cluster samples.

Neter, J., M., H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied Linear Statistical Models. Irwin. A comprehensive text on regression and analysis of variance and covariance.

Olsen, A. R. and E. P. Smith (eds.) 1999. Sampling over time. Special issue of J. of Agric., Biol. and Env. Statistics 4:328-514.

Sauer, J. R., and S. Droege (eds). 1990. Survey designs and statistical methods for the estimation of avian population trends. USFWS Biological Report 90.

Steel, R. G. D. and J. H. Torrie. 1980. Principal Procedures of Statistics. McGraw-Hill. Discusses split-block analysis on pages 390-397.

Thompson, S. K. 1992. Sampling. Wiley. Excellent statistical text having little overlap with Lohr. Covers many useful topics including detectability, line transects, spatial prediction, and adaptive sampling.

Thompson, W.L., G.C. White, and C. Gowan. 1998. Monitoring vertebrate populations. Academic Press, San Diego.

*Estimation of variance components literature:*

Burnham, K. P., D. R. Anderson, G. C. White, C. Brownie, and K. H. Pollock. 1987. Design and analysis methods for fish survival experiments based on release-recapture. American Fisheries Society Monograph Number 5.

Gould, William R., and James D. Nichols, 1998: Estimation of temporal variability of survival in animal populations. *Ecology*: Vol. 79, No. 7, pp. 2531–2538.

Link, W. A., and J. D. Nichols. 1994. On the importance of sampling variance to investigations of temporal variation in animal population size. *Oikos* 69:539–544.

Stewart-Oaten A, Murdoch WW, and Walde SJ. 1995. Estimation of temporal variability in populations. *Am. Nat.* 146:519-35.

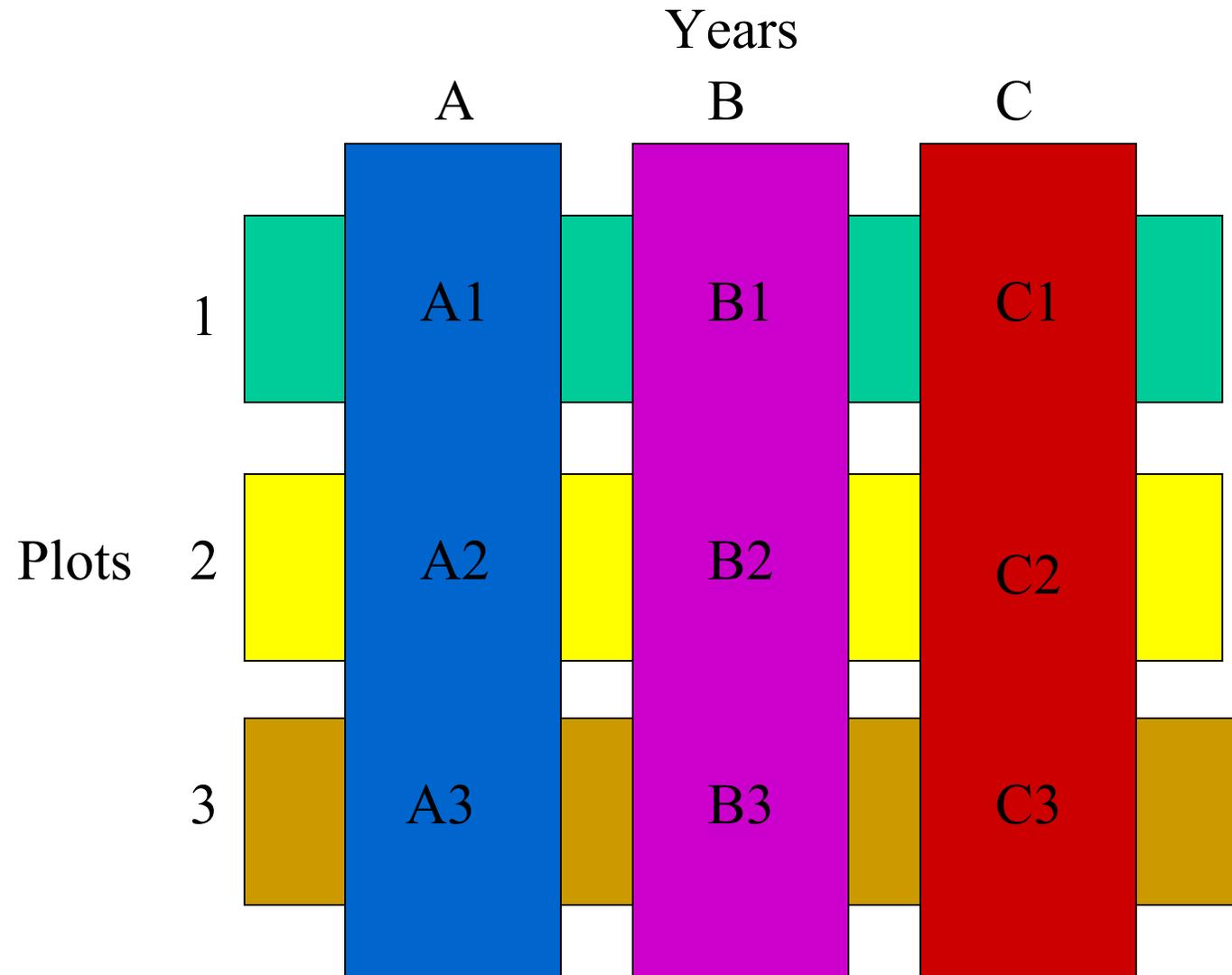
## **Appendix A Workshop Participants**

<u>Name</u>	<u>Affiliation</u>	<u>Phone</u>	<u>Email</u>
Steve Acker	NPS PNR	206 220-4267	Steve_Acker@nps.gov
JeanYves (Pip) Coubois	UW Dept Statistics	206 616-9439	pip@stat.washington.edu
Steve Fradkin	ONP	360 374-1222	Steven_Fradkin@nps.gov
E. Oz Garton	U of Idaho	208 885-7426	ogarton@uidaho.edu
Paul Geissler	USGS-BRD	301 497-5780	Paul_Geissler@usgs.gov
Patti Happe	ONP	360 565-3065	Patti_Happe@nps.gov
Cat Hoffman	ONP	360 565-3060	Cat_Hoffman@nps.gov
Roger Hoffman	ONP	360 565-3062	Roger_Hoffman@nps.gov
Gail Irvine	USGS-BRD	907 786-3653	Gail_Irvine@usgs.gov
Kurt Jenkins	USGS-BRD	360 565-3041	Kurt_Jenkins@nps.gov
Lyman McDonald	West, Inc.	307 634-1256	lmcdonald@west-inc.com
Eric Rexstad	U of Alaska	907 474-7159	ffear@uaf.edu
Regina Rochefort	NOCA	360 856-5700	Regina_Rochefort@nps.gov
Susan Roberts	ONP	360 565-3046	Susan_Roberts@nps.gov
Ed Schreiner	USGS-BRD	360 565-3044	Ed_Schreiner@nps.gov
Andrea Woodward	USGS-BRD	206 526-6282	Andrea_Woodward@usgs.gov

**Appendix B Illustration of Spit-Block Design** – Paul Geissler

The design is like an agricultural field with row and column treatments in strips. The treatments are not randomly assigned to each cell as illustrated below, instead of something like:

B2 C1 A3  
C3 A2 B1  
A1 B3 C2



**Appendix C Example of Channel Islands Data Analysis – Paul Geissler**

Here are results and analysis for cover of exotic vegetation from plots on San Miguel Island in the Channel Islands, California.

Least-square means:

Precipitation	Period 1 ('84-'95)	Period 2 ('96-'00)	Mean
High (>13 in.)	106	102	104
Low (< 13 in.)	81	84	83
Mean	94	93	94

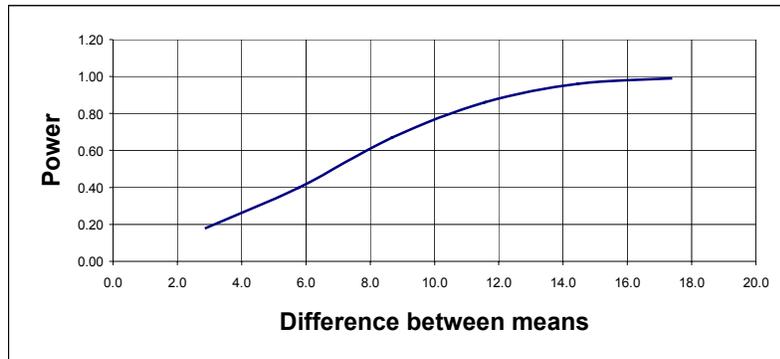
The analysis is for a split-block (a.k.a. split-plot in time) design (Steel and Torrie 1980:390-393)

Analysis of Variance

Source	df	Mean Square	F	P
Site	16	118.780	48.208	0.000
Year	13	23.532	9.551	0.000
Period	1	0.149	0.060	0.806
Precipitation	1	61.423	24.929	0.000
Per. x Precip.	1	2.168	0.880	0.349
Error (= Site x Yr)	180	2.464		

At San Miguel, cover of exotics varied significantly among sites, years and precipitation categories but not with period. Therefore, we conclude that there is no significant trend in exotic cover with time. Results from Santa Rosa Island (not shown here), however, showed a significant response to precipitation, period and precipitation x per, indicating a change in time that depended on precipitation.

Here is a power curve for San Miguel data showing power as a function of differences between means. This curve was generated from the ANOVA using tables from Cohen 1988, Chapter 8. It can also be generated from SYSTAT and the code was provided.



**Appendix D Numeric Example of Calculating Probability of Conformity – Eric Rexstad**

Let's say that after 9 years of measuring a population you estimate:

$$\hat{\mu}_9 = 4.059$$

$$\hat{\sigma}_9 = 0.192$$

In year 10 you observe:

$$y_{10} = 3.674 \text{ (population estimate in year 10)}$$

$$se_{10} = 0.150 \text{ (standard error of population estimate in year 10)}$$

and you want to know the probability that  $y_{10}$  comes from the distribution described by years 1-9.

$$\text{Calculate } t_{10} = \frac{y_{n+1} - \hat{\mu}}{\sqrt{(\hat{\sigma}^2 + se_{n+1}^2)}} = \frac{3.674 - 4.059}{\sqrt{(0.192^2 + 0.150^2)}} = -1.584$$

The probability that the value of  $y$  observed in year 10 belongs to the same distribution as those from years 1-9, named the Probability of Conformity (PC), is:

$$PC_{n+1} = 2p(t_n \geq |t_{n+1}|) = 2p(t_9 \geq |t_{10}|) = 2 \times 0.074 = 0.148 = PC_{10}$$

This means that there is a 15% chance that the observation of  $y$  in year 10 came from the same distribution as years 1-9.

Note: In this example,  $p$  was read from a one-tailed  $t$  table. If you use a two-tailed table you do not need to multiply by 2.

If  $y_{10} = 3.5$  and  $se_{10} = 0.15$ , there is only a 5% probability that the observed  $y$  came from the same distribution as years 1 to 9.

## **Appendix E Comparison of Sample Types—Paul Giessler**

Imagine a sample coming from a gradient with population values of 1 to 9. The true mean is 5. Now take a sample of 3 numbers in three different ways:

### *Simple Random Sample*

If you drew all 84 possible samples, you would calculate an unbiased estimate for the mean (5) and an unbiased estimate of the variance (1.67)

### *Compact Cluster Sample* (bird samples are often taken this way)

Cluster the numbers into threes and take the mean of each cluster:

1 2 3	4 5 6	7 8 9
2	5	8

You would estimate the mean to be 5 and it is unbiased. However, the variance is much larger than for a simple random sample because the groups are autocorrelated. The variance estimate would be 0.22 and underestimates the true variance of 6.00. In an ANOVA, you usually find that the clusters are different. It is possible to calculate a design effect to correct the variance estimate.

### *Systematic Cluster Sample*

1 2 3 4 5 6 7 8 9

A systematic sample of this population would be 1, 4, 7 or 2, 5, 8 or 3, 6, 9. In any case you would have an unbiased estimate of a mean of 5. The variance estimate of 2.00 overestimates the true variance of 0.67. There is no effective technique for correcting the variance. The advantage of this sample scheme is that the true variance is smaller than the variance of a simple random sample.