Standard Operating Procedure (SOP) #23

Statistical Data Analysis

Version 1.01 (June 4, 2021)

Change History

New Version #	Revision Date	Author	Changes Made	Reason for Change	Previous Version #
1.01	6/4/2021	Kim Weisenborn	Converted equations from Equation Editor 3.0, which is no longer supported, to Office Math ML format.	To make equations editable and 508 compliant.	1.0

Only changes in this specific SOP will be logged here. Version numbers increase incrementally by hundredths (e.g., version 1.01, version 1.02) for minor changes. Major revisions should be designated with the next whole number (e.g., version 2.0, 3.0, 4.0). Record the previous version number, date of revision, author of the revision, changes made, and reason for the change along with the new version number.

Purpose

This SOP describes how to analyze Pacific Island Network (PACN) Focal Terrestrial Plant Communities (FTPC) monitoring data for both status and trends. Status is described primary with summary statistics including means and variances, while trends are evaluated using Chi-square tests, McNemar's test of symmetry, paired t-tests, repeated measures ANOVA, or generalized linear models, depending on the type of data and its distribution.

Note that a Type I error level of 0.10 is assumed. Testing at this level may be conservative for long-term monitoring. Since a Type I error occurs when we erroneously find a trend that is not real, the consequences are to take management action to conserve the resource for which the trend was detected. The consequences of a Type II error may be more severe if native plant vigor or recruitment are decreasing without detection. The relative cost of each error make the use of a large Type I error rate reasonable for monitoring (Buhl-Mortensen 1996; Gibbs, et al. 1998; Mapstone 1995).

Status

Based on the certified data, descriptive statistics can be computed for most measured attributes. For continuous variables (e.g., percent cover, density, height, or DBH), summary statistics should include

the mean, standard deviation, skewness, kurtosis, minimum and maximum values. These statistics provide information about the distribution and normality of the dataset. Normal probability plots (or quantile-quantile plots) will also be used to evaluate normality and outliers.

For categorical attributes such as presence/absence, different descriptions are appropriate. For species presence/absence, we will compute three measures of species richness (total, native, and nonnative), as well as a ratio (native richness:total species richness). For substrate type, we will display the data in a vertical bar chart showing the percent of points in each substrate type. Table SOP 23.1 provides a list of recommended summary statistics based on the data collected.

For understory cover, 20% of the fixed plots will be re-read to assess measurement error. A paired t-test will be used to determine if significant measurement errors exist.

Table SOP 23.1. List of recommended summary statistics and trend analysis methods, organized by vegetation attributes. Trend analysis is recommended for variables with asterisks next to them.

Vegetation Attribute	Plot Size (m)	Summary Statistics (Means & Var)	Trend Analysis
Presence/Absence	20 x 50 10 x 20	Total Species Richness* Native Species Richness* Nonnative Species Richness*Ratio (Native Richness:Total Richness)*	Chi-Square (P/A for some species) McNemar's Test (P/A for some species) Generalized Linear Model (Richness)
Understory Cover (<1 m; 1-2 m)	20 x 50 10 x 20	% cover for some species and life forms, native and nonnative (<1 m)* % cover for some species and life forms, native and nonnative (1-2 m)*	Generalized Linear Model*
Understory Cover - QC Tests	20 x 50 10 x 20	Paired t-test	N/A
Large Tree Density (DBH ≥10 cm)	20 x 50	# Trees/ha (all trees)* # Trees/ha (for some species)* # Native Trees/ha* # Nonnative Trees/ha* # Snags/ha* # Native Snags/ha* # Nonnative Snags/ha* # Boles/Tree (all trees) # Boles/Native Tree # Boles/Nonnative Tree DBH (all trees) DBH (for some species)* Percent of Trees with Damage Percent of Natives with Damage Percent of Nonnatives with Damage Percent of Trees Flowering/Fruiting Percent of Natives Flowering/Fruiting Percent of Nonnatives Flowering or Fruiting	Generalized Linear Model*
Tree Canopy Height	20 x 50	Canopy Height	none

Vegetation Attribute	Plot Size (m)	Summary Statistics (Means & Var)	Trend Analysis
Tree Fern Density (Diameter ≥10 cm)	10 x 25	# Tree Ferns/ha* # Tree Fern Snags/ha* Diameter	Generalized Linear Model*
Tree Fern Subcanopy Height	10 x 25	Subcanopy Height	none
Sapling Density	10 x 25	# Saplings/ha (all trees)* # Saplings/ha (for some species)* # Native Saplings/ha* # Nonnative Saplings/ha* # Dead Saplings/ha* # Dead Native Saplings/ha* # Dead Nonnative Saplings/ha* % Saplings with Damage % Native Saplings with Damage % Nonnative Saplings	Generalized Linear Model*
Seedling Density	2 x 50	# Seedlings/ha* # Native Seedlings/ha* # Nonnative Seedlings/ha* % Seedlings in terrestrial substrate* % Native seedlings in terrestrial substrate* % Nonnative seedlings in terrestrial substrate*	Generalized Linear Model*
Shrub Density	2 x 50	# Shrubs/ha* # Native Shrubs/ha* # Nonnative Shrubs/ha* % shrubs in terrestrial substrate* % shrubs at maturity* % Native shrubs in terrestrial substrate % Native shrubs at maturity % Nonnative shrubs in terrestrial substrate % Nonnative shrubs at maturity	Generalized Linear Model*
Tree Fern Juvenile Density	2 x 50	# tree fern juveniles/ha (overall)* # tree fern juveniles/ha (some species)* % Native tree fern juveniles (if possible) % Nonnative tree fern juveniles (if possible) % tree fern juveniles in terrestrial sub. % Native tree fern juveniles in terrestrial substrate	Generalized Linear Model*
Coarse Woody Debris (CWD)	90 m of Transect; 70 m of Transect	# woody logs/ha # tree fern logs/ha Diameter woody logs Diameter tree fern logs % CWD	Generalized Linear Model*

* Generalized Linear Model encompasses repeated measures ANOVA and general linear mixed models.

Formulas for Continuous Data

In most sampling frames (except for mangrove forests) field crews will collect continuous data from both permanent and rotating plots. To compute statistics, we will use data from both plot types. Table SOP 23.2 shows an excerpt of the rotational design from Skalski (2005).

Table SOP 23.2. Rotational design with fixed and rotational sites.

Fixed	Rotational			
	Cycle 1	Cycle 2	. . .	Cycle C
x_{11}	y_{11}	y_{21}		y_{c1}
x_{12}	y_{12}	y_{22}		y_{c2}
x_{13}	⋮	⋮		⋮
x_{14}	y_{1n}	y_{2n}		y_{cn}
⋮				
x_{1k}				

Using the notation above (where $x_{11}, \cdots, x_{1k}$ represents data from fixed sample plots in year 1, $y_{11}, \cdots, y_{1n}$ represents the rotating plots from year 1, etc.), the initial estimate of the population mean in year 1 is

$$\hat{\bar{X}}_1 = \frac{\left(\sum_{i=1}^{k} x_{1i} + \sum_{j=1}^{n} y_{1j}\right)}{(k+n)}, \qquad \textbf{Equation 1}$$

with an estimated variance of

$$\widehat{Var}\left(\hat{\bar{X}}_1\right) = \frac{\left(1 - \frac{(k+n)}{N}\right) s^2_{\text{Pool}}}{k+n} \qquad \textbf{Equation 2}$$

where

$$s^2_{\text{Pool}} = \frac{\sum_{i=1}^{k}\left(x_{1i} - \hat{\bar{x}}_1\right)^2 + \sum_{j=1}^{n}\left(y_{1j} - \hat{\bar{y}}_1\right)^2}{k+n-1} \qquad \textbf{Equation 3}$$

and

$$\left(1 - \frac{(k+n)}{N}\right) \qquad \textbf{Equation 4}$$

represents the finite population correction (FPC) factor (Skalski 2005). Note that the FPC factor is likely to equal one and probably can be ignored. In the second year of sampling the initial estimate of the population mean would be:

$$\hat{\bar{X}}_2 = \frac{\left(\sum_{i=1}^{k} x_{2i} + \sum_{j=1}^{n} y_{2j}\right)}{(k+n)}. \qquad \textbf{Equation 5}$$

According to Skalski (2005) the repeated measures from year 1 to year 2 can be used to improve the precision of the prior year's estimate (i.e., $\hat{\bar{X}}_1$) by regressing x_{1i} on x_{2i} (i.e., $x_{1i} = \alpha + \beta x_{2i} + \varepsilon_i$) to provide an alternative year 1 estimate:

$$\hat{\bar{X}}_1' = \hat{\alpha} + \hat{\beta}\left(\frac{\sum_{i=1}^{k} x_{2i}}{k}\right) = \hat{\alpha} + \hat{\beta}\bar{x}_2. \qquad \textbf{Equation 6}$$

With the two estimates for year 1 ($\hat{\bar{X}}_1$ and $\hat{\bar{X}}_1'$), the best year 1 estimate is a variance weighted estimate using the following Skalski (2005) formula:

$$\tilde{X}_1 = \frac{\frac{1}{\mathrm{Var}\left(\hat{\bar{X}}_1\right)}\hat{\bar{X}}_1 + \frac{1}{\mathrm{Var}\left(\hat{\bar{X}}_1'\right)}\hat{\bar{X}}_1'}{\frac{1}{\mathrm{Var}\left(\hat{\bar{X}}_1\right)} + \frac{1}{\mathrm{Var}\left(\hat{\bar{X}}_1'\right)}}. \qquad \textbf{Equation 7}$$

As sampling continues, each year we can compute the current year's estimate, plus an adjusted estimate for the prior year.

Formulas for Discrete Data

This section applies to presence/absence (P/A) data and percent cover data in the two understory layers. Using discrete P/A values, species richness will be tabulated as the count (or number) of different species found in a given plot. Species richness is subdivided into native and nonnative species richness, and also expressed as the ratio of native species richness to total species richness. The terms below summarize this information:

N_{total} = total species richness (count of total species in a plot),

n_{native} = native species richness (count of native species in a plot),

$n_{nonnative}$ = nonnative species richness (count of nonnative species in a plot), and

$\frac{n_{native}}{N_{total}}$ = ratio of native species richness to total species richness.

For the two understory layers (<1 m and 1-2 m), percent cover will be computed using the point-intercept method for 300 points per plot. Based on the number of "hits" observed for a given species (or group of species, e.g., natives), percent cover will be computed as the number of "hits" observed divided by total number of "hits" possible, as described below:

$$Percent\ Cover_{species\ X} = 100\left(\frac{n_{observed}}{N_{possible}}\right), \qquad \textbf{Equation 8}$$

where $n_{observed}$ = the number of "hits" observed, and

$N_{possible} = 300.$

Aggregating Data from Multiple Sampling Frames

For plant communities with more than one sampling frame (e.g., the wet forest of HAVO and NPSA), scientists and managers often want to see the data aggregated for the entire plant community, not just a particular sampling frame within that community. In this case, the standard formulas for stratified random sampling apply where each sampling frame represents a different stratum within the plant community. For instance, data from the three wet forest sampling frames at HAVO can be combined to create overall statistics for the entire wet forest community. Following from Skalski (2005), the formula to compute the overall population mean from strata (or analogously, the plant community mean from multiple sampling frames) is:

$$\hat{\bar{X}} = \frac{\sum_{g=1}^{L} \hat{\bar{X}}_g \cdot A_g}{\sum_{g=1}^{L} A_g} = \sum_{g=1}^{L} \hat{\bar{X}}_g W_g \qquad \textbf{Equation 9}$$

where $\hat{\bar{X}}_g$ = estimate of the g^{th} strata mean (expressed in consistent areal units such as m^2, km^2)

A_g = area of the g^{th} stratum,

L = number of strata in the sampling frame, and

$W_g = \frac{A_g}{\sum_{g-1}^{L} A_g}$ = weight of the g^{th} stratum.

The variance of $\hat{\bar{X}}$ is

$$\text{Var}\left(\hat{\bar{X}}\right) = \sum_{g=1}^{L} W_g^2 \cdot \text{Var}\left(\hat{\bar{X}}_g\right). \qquad \textbf{Equation 10}$$

Trends

In the framework of the rotating panel design, only monitoring data from permanent plots will be used for trend analysis. Of the attributes listed in Table SOP 23.1, the ones with asterisks next to them will be analyzed for trend. The list of continuous variables includes plant community composition variables such as richness and structural variables such as percent cover and density (by layer, plant life form, size, and maturity level). Trends in categorical data (presence/absence) will be analyzed using the Chi-square goodness of fit test and McNemar's test of symmetry. These tests evaluate whether the proportion of plots with species presence changes between sampling periods. Based on initial monitoring data, the project lead will choose which species are appropriate for Chi-square and McNemar analysis.

Once two sampling periods of data exist for continuous attributes, parametric paired t-tests can be used to assess changes between sampling periods for continuous variables that meet the standard assumptions of normality and homogeneity of variance. Note that in terms of normality, it is the normality of residuals (not the normality of the raw data) that is required for significance testing

(Kery and Hatfield 2003). After three or more sampling periods of data exist, repeated measures ANOVA can be used provided the standard assumptions are met. A more generally applicable method for assessing the effect of year on a response variable is to use a generalized linear model (GzLM) that can accommodate a variety of error structures.

Generalized Linear Model

Following from Schneider (2007) an initial GzLM model can be written as

$$N = \mu + Normal\ error \qquad \textbf{Equation 11}$$

$$\mu = \beta_o + \beta_P \cdot P + \beta_{Yr} \cdot Yr \qquad \textbf{Equation 12}$$

where N = response variable (e.g., seedling density),

Yr = year (the categorical explanatory variable of interest), and

P = plot (the secondary explanatory variable for statistical control, also categorical).

Substituting the second equation into the first, we obtain the traditional general linear model (in this case, a two-way ANOVA) with normal error distribution. For data that meet the standard assumptions of normality and homogeneity of errors, the above model is adequate. However, when standard ANOVA assumptions are not met, the generalized linear model allows us to move from normal errors to other error distributions (e.g., negative binomial, gamma, etc.). A brief example follows (Schneider 2007).

Hypothetical response variable data for two years is presented in Table SOP 23.3, while the same data reorganized for input into a GzLM routine is presented in Table SOP 23.4. Both SAS and SPlus require data in the format shown in Table SOP 23.4. Other statistical packages may have different input requirements.

Table SOP 23.3. Sample response variable data for 20 plots after two years of monitoring.

Plot	Year 1	Year 2
1	222	171
2	125	101
3	69	57
4	92	161
5	121	792
6	97	121
7	153	119
8	609	360
9	147	93
10	135	130
11	32	73
12	113	200
13	51	48
14	62	92
15	78	105
16	26	35
17	10	6
18	5	10
19	59	62
20	87	48

Table SOP 23.4. Reorganized response variable data for 20 plots after two years of monitoring.

Plot	Year	N
1	1	222
2	1	125
3	1	69
4	1	92
…	…	…
18	1	5
19	1	59
20	1	87
1	2	171
2	2	101
3	2	57
4	2	161
…	…	…
18	2	10
19	2	62
20	2	48

```
******************** GzLM with normal error structure *****************;

proc genmod data = Input_Data;
      class plot year;
      model N = plot year /
            link = identity
            dist = normal
            scale = deviance
            type1 type3;
OUTPUT out=RESPRED p=pred resraw=resraw;
run;

proc gplot data=RESPRED;            *plot residuals vs. predicted;
plot resraw*pred;
run;

proc univariate data = RESPRED normal plot;     *QQ plot, normality tests;
var resraw;
run;

******************** GzLM with gamma error structure ***********************;

proc genmod data = Input_data;
      class plot year;
      model N = plot year /
            link = identity
            dist = gamma
            scale = deviance
            type1 type3;
OUTPUT out=RESPRED2 p=pred resdev=resdev;
run;

proc gplot data=RESPRED2;           *plot deviance residuals vs. predicted;
plot resdev*pred;
run;
```

Figure SOP 23.1. Sample SAS code to run a GzLM on the data in Table SOP 23.3.

Based on the hypothetical data in Tables SOP 23.3 and 4, Figure SOP 23.1 contains the SAS commands required to run two different models, one with normal error structure and another with gamma error structure. The code also performs diagnostic tests on each model to evaluate its appropriateness. Since year has only two classes in our sample data, the first model represents a paired comparison design comparable to a paired t-test. The diagnostic tests on this model indicate that the residuals are neither normal (fig. 2), nor exhibit homogeneity of errors (fig. 3). Consequently, since the standard assumptions are violated, we cannot trust the p-values from this model.

The second model run in Figure SOP 23.1 uses a gamma distribution to model the error structure, expressed in Equations 11 and 12. As a diagnostic test, scaled residuals, or deviance residuals, are plotted against the fitted values. In models with non-normal error structures, scaled residuals must be plotted instead of raw residuals. In the second model the deviance residuals are homogenous, the

assumptions met, and the p-values reliable (Schneider 2007). Schneider provides additional details on fitting models with different error structures and links (e.g., log links examine multiplicative effects while identity links examine additive effects).

Since the assumptions are met in this case, we can evaluate the null hypothesis ($H_o: \beta_{Yr_1} = \beta_{Yr_2}$, or alternatively expressed, H_o: Deviance $(\beta_{Yr}) = 0$) based on the analysis of deviance table. In SAS, either Type 1 or Type 3 analysis (more commonly used) can be used to evaluate the potential year effect on the response variable. Type 1 analysis provides a Chi-Square value for adding a year effect to all previous effects (intercept and plot), while Type 3 analysis evaluates the effect of removing the year effect from the complete model. In this case, the two types of analysis are identical (fig. 4). Schneider (2007) provides a more detailed interpretation of the analysis of deviance table and the GLzM output within the context of SPlus. Essentially Schneider recommends using a G-statistic to evaluate year effect, which in SAS terms corresponds to the Chi-square value. The only difference between Schneider's G-statistic and the SAS Chi-square value is that the SAS Chi-square is adjusted for the model's dispersion factor. In this example, the Chi-square value of 0.29 does not allow us to reject the null hypothesis for a year effect; consequently we must conclude that no significant year effect exists.

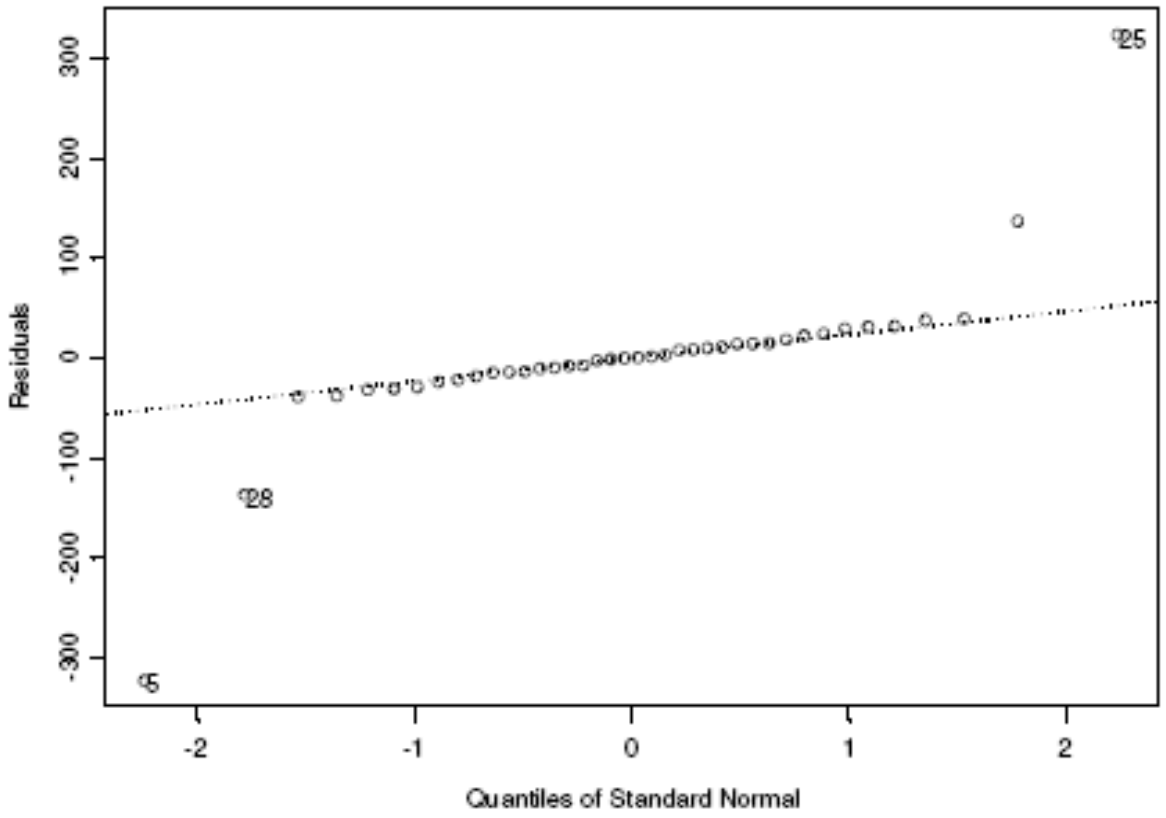

Figure SOP 23.2. QQ-plot of residuals (from SPlus) for initial model with normal error structure. *Source:* Schneider 2007. Note that the residuals do not fall on a straight line and are not normal.

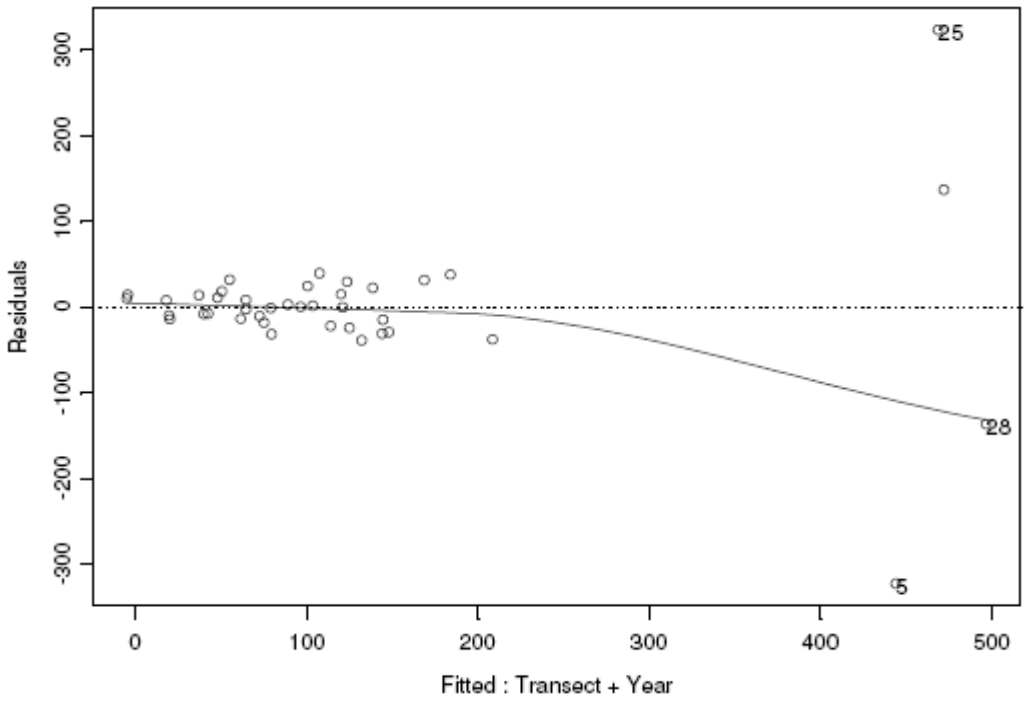

Figure SOP 23.3. Plot of residuals against fitted values (from SPlus), showing increasing variance with larger fitted values. *Source:* Schneider 2007.

LR Statistics For Type 1 Analysis

Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	38.0509						
Plot	3.2793	19	19	10.77	<.0001	204.54	<.0001
Year	3.2300	1	19	0.29	0.5964	0.29	0.5901

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Plot	19	19	10.66	<.0001	202.62	<.0001
Year	1	19	0.29	0.5964	0.29	0.5901

Figure SOP 23.4. Analysis of deviance output, Type 1 analysis, and Type 3 analysis (most commonly used) for the model with gamma error, from SAS.

Literature Cited

Buhl-Mortensen, L. 1996. Type-II statistical errors in environmental science and the precautionary principle. Marine Pollution Bulletin 32:528-531.

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

Kery, M. and J. S. Hatfield. 2003. Normality of Raw Data in General Linear Models: the Most Widespread Myth in Statistics. Bulletin of the Ecological Society of America 84:92-94.

Mapstone, B.D. 1995. Scalable decision rules for environmental impact studies: effect size, Type I, and Type II errors. Ecological Applications 5:401-410.

Schneider, D. 2007. Example of Generalized Model (GzLM) Using Splus. National Park Service Pacific Island Network Unpublished Report, Hawaii National Park, Hawaii.

Skalski, J. R. 2005. Long-term monitoring: Basic study designs, estimators, and precision and power calculations. National Park Service Pacific Island Network Unpublished Report, Hawaii National Park, Hawaii.