Standard Operating Procedure (SOP) #21

Statistical Data Analysis

Version 1.01 (May 26, 2021)

Change History

New Version #	Revision Date	Author	Changes Made	Reason for Change	Previous Version #
1.01	5/26/2021	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 established invasive plant species monitoring data for both status and trends for Pacific Island Network (PACN) parks. Status is described primarily with summary statistics including means and variances, while trends are evaluated using paired t-tests, repeated measures analysis of variance (ANOVA), generalized linear models, zero-inflated generalized models, proportional odds models, or likelihood ratio tests depending on the type of data and its distribution. A list of recommended summary and trend statistics based on the data collected is provided in Table SOP 21.1.

Table SOP 21.1. List of recommended summary statistics and trend analysis methods for each vegetation attribute.

Vegetation Attribute	Summary Statistics (Means & Var)	Trend Analysis
Invasive Species Richness	Count of invasive species per plot	Generalized Linear Model Repeated Measures ANOVA Paired t-test
Frequency	No. or % of plots with presence of species, life forms, or other groupings	Zero-Inflated Generalized Model Likelihood Ratio Test
Cover Class	No. or % of plots with a particular cover class for a species	Proportional Odds Model (Ordered Logit) Likelihood Ratio Test

Status

Based on the certified presence/absence data, descriptive statistics (means, variances, confidence intervals, etc.) can be computed for target species frequency and invasive species richness with the transect as the sample unit. Invasive species richness per transect will be calculated as the average count (or number) of different species found in each contiguous plot along the transect. The sampling design is a one-stage cluster sample, where the primary sampling unit is a transect and the secondary sampling unit is a plot. Species frequency will be calculated as the proportion of sampled plots along a transect where a species or specific group of species are found. Cover class data can also be used to compute descriptive statistics at the transect level. A separate status estimate will be reported for each cover class for each species with their associated error. Depending on the attribute, these statistics are aggregated across all species, grouped by life form (i.e., tree, shrub, fern, herbaceous), and/or individual species.

There are two design-based estimators available for status summary with a one-stage cluster sample with unequal transect lengths (primary sampling unit size varies), a ratio estimator and an unbiased estimator. Ratio estimates of the population average per plot are based on the ratio of the total of all the values for the variable in all the plots in the sample and the total number of plots in the sample. Unbiased estimates of the population average per plot are based on the average for all transects sampled, where the observation is the average value per plot for a transect, and can better address transects of differing lengths. If the transects are all of equal length, the ratio and unbiased estimator are equivalent; however, if the transect lengths vary and the number of plots with a species is proportional to the length of the transect the ratio estimator will be more efficient (smaller variance). In other words, in the situation that longer transects have more plots with presences recorded, the ratio estimator would be preferred.

For the equations below, we used the following notation:

N = the number of primary sampling units in the population;
n = the number of primary sampling units in the sample;
i indexes the transect where $i = 1,\ldots,N$;

M_i = the number of plots within the *i*th transect;
j indexes the plots within a transect where $j = 1,\ldots, M_i$;
M = $\sum_{i=1}^{N} M_i$ is the number of secondary sampling units in the population;
$\bar{M}$ = the average PSU size in the population;
y_{ij} = the observation recorded in plot *j* in transect *i*; for our report this is the plot invasive species richness, the presence or absence of a species within a plot, or the presence of a species within a particular cover class.
$y_i = \sum_j y_{ij}/M_i$ is the average invasive species richness per plot for transect *i*, the proportion of plots occupied by a species in transect *i*, or the proportion of plots occupied by a species within a particular cover class.

Ratio Estimator

For a ratio estimator of the population average per plot ($\hat{\bar{y}}_r$), we use (Lohr 2010)

$$\hat{\bar{y}}_r = \frac{\sum_i M_i \bar{y}_{\imath}}{\sum_i M_i} \qquad \textbf{Equation 1}$$

The variance estimator is (Lohr 2010)

$$Var(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right)\left(\frac{1}{n\bar{M}^2}\right)\left(\frac{\sum_i {M_i}^2 \left(\bar{y}_i - \hat{\bar{y}}_r\right)^2}{n-1}\right) \qquad \textbf{Equation 2}$$

Unbiased Estimator

If the area of the sample frame is known, the unbiased estimator for the population average per plot (Thompson 2002) is

$$\hat{\bar{y}}_{unb} = \frac{\hat{\tau}}{M}\,, \qquad \textbf{Equation 3}$$

where

$$\hat{\tau} = \frac{N}{n} \sum_i M_i \bar{y}_i\ . \qquad \textbf{Equation 4}$$

The variance estimator is (Thompson 2002)

$$\widehat{Var}(\hat{\bar{y}}_{unb}) = \frac{s^2}{n} \qquad \textbf{Equation 5}$$

where

$$s^2 = \frac{1}{n-1} \sum_i \left(\bar{y}_i - \bar{\hat{y}}_{unb}\right)^2. \qquad \textbf{Equation 6}$$

Aggregating Data

For areas or communities with more than one sampling frame (e.g., the wet forest of HAVO and NPSA), we may want to aggregate the data for the entire area, not just a particular sampling frame. In this case, the standard formulas for stratified random sampling apply where each sampling frame represents a different stratum within the plant community. Following from Skalski (2005), the formula to estimate the overall population mean from strata (or analogously, the invasive species mean from multiple sampling frames) is:

$$\hat{\hat{X}} = \frac{\sum_{g=1}^{L} \hat{\hat{X}}_g \cdot A_g}{\sum_{g=1}^{L} A_g} = \sum_{g=1}^{L} \hat{\hat{X}}_g W_g$$ **Equation 7**

where $\hat{\hat{X}}_g$ = estimate of the g^{th} strata (mean species richness or frequency [sample proportion])
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 $\hat{\hat{X}}$ is

$$\mathrm{Var}\left(\hat{\hat{X}}\right) = \sum_{g=1}^{L} W_g^2 \cdot \mathrm{Var}\left(\hat{\hat{X}}_g\right)$$ **Equation 8**

Trends

For invasive richness, parametric paired t-tests are used to assess changes between the first two sampling periods if data meet the standard assumptions of normality and homogeneity of variance. It is the normality of residuals (not the normality of the raw data) that is required for significance testing (Kery and Hatfield 2003). For a paired t-test, it is the differences that are assumed consistent with a normal distribution. After three or more years of richness data exist, repeated measures ANOVA are used provided the standard assumptions are met. Alternatively, 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 (Schneider 2007).

For both species frequency and species cover class data, the general process is to fit two models that differ only in the inclusion (full model) or exclusion (reduced model) of year to the data and compare the output using the likelihood ratio test. A significant p-value indicates that there is a significant change in the variable between years. For frequency data, the models used are logit models (a non-lineal mixed model) with a zero-inflated beta distribution. For species cover class data, proportional odd models are used instead. For all variables, trends that yield a p-value of less than 0.1 ($p < 0.1$) are deemed significant for our purposes. Based on initial monitoring data, the project lead will choose which species or groups of species are appropriate for analysis.

Species Richness

Since we are interested in transect level richness, we first average the plot richness values for each transect. The central limit theorem implies that the transect average richnesses will be approximately normal because they are based on averages of plot level species richness. Then, we use mixed model ANOVA to test for differences over years with transects being random and years being fixed. The underlying model is

$$\gamma_{ij} = \mu + Tr_i + \gamma_j + e_{ij} \qquad \textbf{Equation 9}$$

where y_{ij} is the average richness in year *j* for transect i,
μ is the overall mean,
Tr_i is the effect of the i-th randomly selected transect,
γ_j is the effect of the j-th year, and
e_{ij} is residual error.
Note that $Tr_i \sim N(0, \sigma_{tr}^2)\ and\ e_{ij} \sim N(0, \sigma^2)$

The hypotheses of interest are $H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_T \quad H_1 : some\, \gamma_j \neq \gamma_{j'}$, To demonstrate the analysis of variance (ANOVA) method, we used data from Ainsworth et al. (2008), a pilot study employing five transects with 5 x 50 m contiguous plots that were surveyed for nonnative species presence in 2000 and in 2008. We could also have used a paired t-test since there were only 2 years of data. The average richness data for each transect and year are presented in Table SOP 21.2 and the output from the ANOVA is presented in Figure SOP 21.1. The average nonnative species richness per plot in 2008 is estimated as 1.6 which is 0.64 species greater than in 2000 (p = 0.0589). In this study the variation among transects is 1.44, 10 times the residual variance of 0.15. This indicates that adding transects would improve the sampling design to better represent the population; the Established Invasive Plant Species Monitoring Protocol has four times as many transects. These data demonstrate how we can determine if target nonnative species richness has changed over time in each proposed sample frame.

Table SOP 21.2. Average target invasive species richness for Ainsworth et al. (2008).

Observation	Year	Transect	Average Richness
1	2000	1	0.70
2	2000	2	0
3	2000	3	0.57
4	2000	4	1.11
5	2000	5	2.58
6	2008	1	1.20
7	2008	2	0.38
8	2008	3	0.78
9	2008	4	1.61
10	2008	5	4.17

```
                     The Mixed Procedure for Species Richness

                 Class       Levels    Values
                 Year             2    2000 2008
                 Transect         5    1 2 3 4 5

                            Cov Parm     Estimate
                            Transect       1.4416
                            Residual       0.1482

                          Solution for Fixed Effects
                                        Standard
      Effect       Year            Estimate       Error      DF    t Value    Pr > |t|
      Intercept                      1.6332      0.5639       4       2.90      0.0443
      Year                 2000     -0.6375      0.2435       4      -2.62      0.0589
      Year                 2008           0           .       .          .          .

                        Type 3 Tests of Fixed Effects
                              Num     Den
                   Effect      DF      DF    F Value    Pr > F
                   Year         1       4       6.86    0.0589
```

Figure SOP 21.1. Mixed-model analysis of variance for change in species richness over time for Ainsworth et al. 2008. The change in mean is significant (p = 0.059).

When transects from different sampling frames are analyzed together (e.g., multiple wet forest strata within NPSA), we include a blocking (stratification) term for this effect in the model,

$$y_{ij} = \mu + \alpha_i + Tr_{j(i)} + \gamma_k + e_{ijk}$$

Equation 10

where αi is the blocking effect of the i-th area and the subscript j(i) indicates that transects are nested within areas.

Data from a nonnative species transect based pilot study by Jacobi and Bio (2001) demonstrate how to combine analysis over sampling areas within a single community type into an overall test of trend. For this study, three wet forest areas were sampled in 1999, 2000, and 2001 along transects with 3 x 10 m contiguous plots. Nonnative richness was similar across the study areas

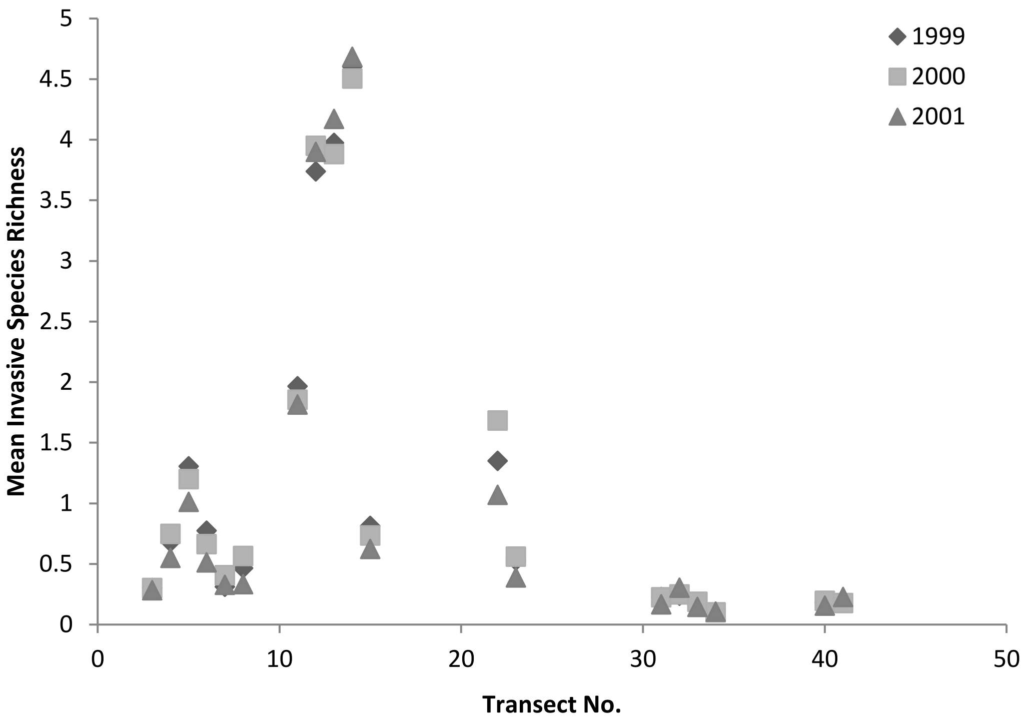
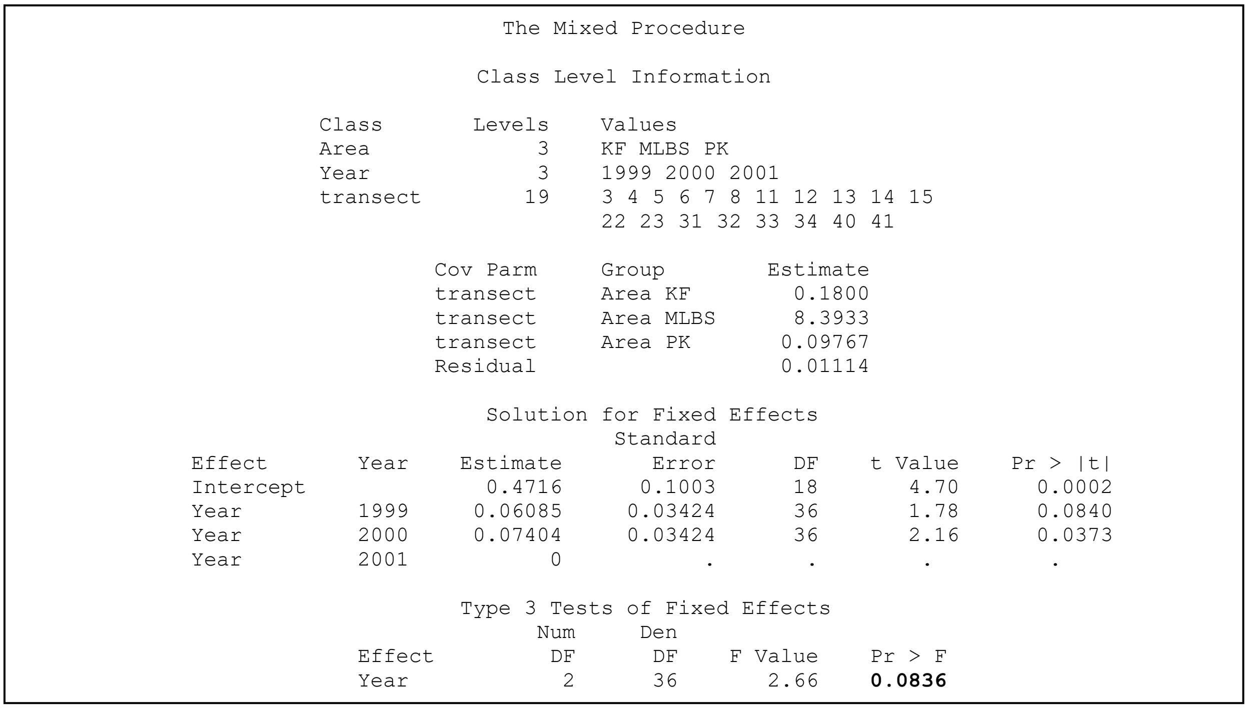


Figure SOP 21.2. Average nonnative species richness for Jacobi and Bio (2001). Transects 3-8 were located in the Mauna Loa Boys' School, 11-15 in Pu'u Kipu, and 22-41 in Kīlauea Forest.



```
                              The Mixed Procedure

                            Class Level Information

                Class        Levels    Values
                Area              3    KF MLBS PK
                Year              3    1999 2000 2001
                transect         19    3 4 5 6 7 8 11 12 13 14 15
                                       22 23 31 32 33 34 40 41

                         Cov Parm      Group        Estimate
                         transect      Area KF        0.1800
                         transect      Area MLBS      8.3933
                         transect      Area PK       0.09767
                         Residual                    0.01114

                             Solution for Fixed Effects
                                       Standard
      Effect       Year    Estimate       Error      DF    t Value    Pr > |t|
      Intercept              0.4716      0.1003      18       4.70      0.0002
      Year         1999     0.06085     0.03424      36       1.78      0.0840
      Year         2000     0.07404     0.03424      36       2.16      0.0373
      Year         2001           0           .       .          .           .

                           Type 3 Tests of Fixed Effects
                                 Num     Den
                   Effect         DF      DF    F Value    Pr > F
                   Year            2      36       2.66    0.0836
```

Figure SOP 21.3. Mixed-model analysis of variance for change in species richness over time in three study areas for Jacobi and Bio 2001. The change in mean between 1999 and 2001 as well as 2000 and 2001 are significant (p = 0.08, p = 0.037 respectively).

except for three of the five transects at Mauna Loa Boys' School (MLBS) (fig. 2). It is important to stratify (or block) these data by area to accommodate for the observed variability and still test for an overall trend. A mixed model analysis of variance for the 57 average richnesses shows the changes in estimated average richness between 1999 and 2001 as well as 2000 and 2001 are significant (p = 0.08, p = 0.037 respectively). The estimated average richness for 2001 was 0.47 species which is slightly lower than in 2000 (0.53 species) and 1999 (0.55 species). Note that the ANOVA output (fig. 3) includes four variance components – one for each area plus the residual. As expected, the variance for MLBS is high supporting the decision to stratify the data by area for analyses. These data demonstrate how we can determine if target nonnative species richness has changed over time for multiple sampling frames across a community type.

Frequency

To address the non-normality of the frequency data due to many values close to zero, we plan to fit models using a zero-inflated beta (0-beta) distribution which is a mixture of a point mass at zero and a beta random variable with range 0<p<1. The parameters of this distribution include p_0, the proportion of zeros, and the parameters of the beta distribution. We take the beta parameters to be its mean, μ, and scale parameter, ϕ. We then use a logit model (a non-linear mixed model) to fit the data for p_0 and μ. That is, we assume that

$$\log\left(\frac{p_{0ij}}{1-p_{0ij}}\right) = \gamma_0 + tr_i + \gamma_j \qquad \textbf{Equation 11}$$

where γ_0 is the intercept
tr_i is the random effect of the i-th transect, and
γ_j is the effect of year j.

Also,

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + tr_i + \beta_j \qquad \textbf{Equation 12}$$

that is, the mean of p>0 has a logit-linear model in transect and year effects.

Under these two models, our null hypothesis is

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_T \; and \; \beta_1 = \beta_2 = \cdots = \beta_T \qquad \textbf{Equation 13}$$

where T is the number of years sampled.

The null hypothesis states that year does not affect the proportion of zeros or the mean of the non-zeros. Alternatively, if we are interested in a linear increase or decrease over years, we would use the models

$$\log\left(\frac{p_{0ij}}{1-p_{0ij}}\right) = \gamma_0 + tr_i + \gamma_1 t$$

Equation 14

and

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_0 + tr_i + \beta_1 t$$

Equation 15

where t denotes the year.

The null hypothesis states that there is not a linear increase or decrease in either the proportion of zero plots or the mean proportion among nonzero plots for a given species

$$H_0\text{: } \gamma_1=0 \text{ and } \beta_1=0.$$

Equation 16

To combine data from multiple areas or sampling frames, we must account for the areas in the model. To do this, we added a sample area term yielding

$$\log\left(\frac{p_{0ijk}}{1-p_{0ijk}}\right) = \gamma_0 + \alpha_i + tr_{j(i)} + \gamma_k$$

Equation 17

and

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \beta_0 + \delta_i + tr_{j(i)} + \beta_k.$$

Equation 18

Using frequency data for *Psidium cattleianum* (PSICAT) from Ainsworth et al. (2008) (Table SOP 19.3), we demonstrate one way to test the null hypothesis that frequency or the proportion occupied by PSICAT did not change over time. First using a 0-beta distribution, we fit two separate non-linear mixed models that differed only in the inclusion (full model) or exclusion (reduced model) of the variable for year and computed the log likelihoods for each model (-0.33 and 2.7 respectively). Then using the likelihood ratio test, we compared the log-likelihoods of the two models. In this case, there was no significant difference between the models (p=0.21), and therefore PSICAT did not significantly change between years.

Table SOP 21.3. Frequency or proportion of plots containing *Psidium cattleianum* (PSICAT) from Ainsworth et al. (2008).

Transect	% PSICAT	
	2000	2008
1	0.100	0.400
2	0.000	0.222
3	0.053	0.053
4	0.056	0.333
5	0.412	0.941

Cover Class

Because the cover class data is categorical instead of continuous, we must use a different model to fit the data. Our preferred model is the proportional odds model also known as the ordered logit model. For the full model, we assume

$$\log\left(\frac{p_{kij}}{1-p_{kij}}\right) = \gamma_{k0} + tr_i + \gamma_{\mathrm{j}} \qquad \textbf{Equation 19}$$

where p_k represents the probability of category *k,*
γ_0 is the intercept,
tr_i is the random effect of the *i*-th transect, and
γ_{j} is the effect of year *j*.

As for the frequency trend analysis, we also fit a reduced model with year excluded and then compare the models using the likelihood ratio test.

Literature Cited

Ainsworth, A., B. Stevens, L. Hadway, N. Agorastos, I. Cole, and C. M. Litton. 2008. Vegetation response to eight years of feral pig (*Sus scrofa*) removal in Pu‘u Maka‘ala Natural Area Reserve, Hawai‘i. State of Hawaii, Division of Forestry and Wildlife.

Jacobi, J. D. and K. Bio. 2001. Invasive Plant Species Surveys, Olaa-Kilauea Management Area. Department of the Interior, US Geological Survey, Biological Resources Discipline, Kilauea Field Station, Hawaii National Park, HI. Unpublished data.

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.

Lohr, S. L. 2010. Sampling: Design and Analysis. Brooks/Cole, Cengate Learning, Boston, MA.

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

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

Thompson, S. K. 2002. Sampling. John Wiley and Sons, New York.