USGS
science for a changing world

Prepared in cooperation with U.S. Fish and Wildlife Service

Sample Size Estimation for Savanna Monitoring Protocol Development

Open-File Report 2022–1053

U.S. Department of the Interior
U.S. Geological Survey

Sample Size Estimation for Savanna Monitoring Protocol Development

By Deborah A. Buhl

Prepared in cooperation with U.S. Fish and Wildlife Service

Open-File Report 2022–1053

U.S. Department of the Interior
U.S. Geological Survey

U.S. Geological Survey, Reston, Virginia: 2022

Suggested citation:
Buhl, D.A., 2022, Sample size estimation for savanna monitoring protocol development: U.S. Geological Survey Open-File Report 2022–1053, 49 p., https://doi.org/10.3133/ofr20221053.

ISSN 2331-1258 (online)

Acknowledgments

These analyses were done in cooperation with the Savanna Monitoring Protocol Team: Pauline Drobney (U.S. Fish and Wildlife Service), Daniel Dey (U.S. Forest Service), and Diane Larson (U.S. Geological Service [USGS]). I thank Amy Symstad (USGS) and Karen Ryberg (USGS) for reviewing earlier drafts of this report.

Contents

Figures

Tables

Conversion Factors

International System of Units to U.S. customary units

Multiply	By	To obtain
	Length	
centimeter (cm)	0.3937	inch (in)
meter (m)	3.281	feet (ft)
	Basal area	
square meters per hectare (m^2/ha)	4.356	square feet per acre (ft^2/acre)

Abbreviations

CV	coefficient of variation
DBH	diameter at breast height
GLM	generalized linear model
GLMM	generalized linear mixed model
n	number of sample plots
NWR	National Wildlife Refuge
p	binomial or multinomial probability

Sample Size Estimation for Savanna Monitoring Protocol Development

By Deborah A. Buhl

Abstract

When designing data collection protocols for a new research project, it is important to have a large enough sample size to detect a desired effect, but not so large as to waste time collecting more data than are needed. Power analysis methods can be used to estimate this sample size. In this report, power analyses used to estimate sample sizes needed for a savanna monitoring study, for which the U.S. Fish and Wildlife Service are developing protocols, are described. Power analyses were run to estimate the sample sizes needed to detect a specified difference (that is, effect size) between means from two savanna areas or between yearly means for a savanna area. Sample sizes were estimated for nine different vegetation metrics that will be measured in savanna areas. Analyses were run for each metric using a range of means and variances, effect sizes, and correlation among repeated measures. Sample size estimates varied among vegetation metrics. Within each vegetation metric, estimated sample sizes varied with means, variances, effect size, and correlation. Many of the sample size estimates were too large to be feasible when sampling; therefore, the tables of estimated sample sizes may be first used as a guide to determine an adequate and feasible sample size that will detect differences in some vegetation metrics. Then, using this sample size, the tables can be used to estimate the effect sizes for each vegetation metric that may be detectable for a given mean, variance, and correlation.

Introduction

Power analysis (Steidl and others, 1997) can be used to estimate the sample size needed to detect a specified difference (that is, effect size) at a given statistical significance level. The goal of sample size estimation is to estimate a sample size that will be large enough to detect the desired effect, but not so large that resources are wasted by collecting more data than are needed. The sample size needed in a study can be affected by factors such as desired effect size, mean and variance of data, significance level, statistical power, experimental design, distribution of data, statistical model, and correlation among repeated measures.

Several steps are involved in doing a power analysis to estimate sample size. First, determine the statistical models that will be used to address the questions of the study after the data have been collected. Part of determining the correct model is identifying the correct distribution of each of the response variables (that is, variables of interest). Second, estimates of the expected means and variances for each variable are needed. These can be determined from literature, existing data, or from a pilot study. Next, an effect size needs to be defined for each variable; the effect size is the difference between mean values that could be significantly detected. Then for desired significance and statistical power levels, the above information can be used to determine the sample size needed to detect the desired effect size. For some statistical models and data distributions, there are equations that can be used to estimate this sample size. For other models and distributions, equations are not readily available; therefore, simulation methods can be used to estimate the sample size.

This report summarizes the power analyses done to aid in survey protocol development for the savanna monitoring study by the U.S. Fish and Wildlife Service. Power analyses were used to estimate the sample sizes needed to detect a specified difference (that is, effect size) between two means. Sample sizes were estimated for two different objectives using nine different vegetation metrics that will be measured in savanna areas (appendix 1).

Savanna Monitoring Study Design and Questions

In the savanna monitoring study, vegetation within savanna areas will be monitored to assess the current state (or health) of the savanna and changes through time. Within this report, a savanna area is defined as a discrete area of savanna habitat. Vegetation data will be collected from a series of plots within each savanna area. Tree macroplots are the largest sized plots and will be used for adult trees (greater than [>] 12.7 centimeters [cm] diameter at breast height [DBH]) and overstory metrics. Trees to be sampled will be identified using a prism sampling method. Nested frequency plots (DeBacker and others, 2011) will be used to measure herbaceous species composition in terms of relative frequency. Nested frequency

plots are a set of nested subplots with the smallest subplot 12.5 by 50 cm and largest 0.5 by 2 meters (m). Saplings (2.5–12.7 cm DBH), oak seedlings (less than [<] 2.5 cm DBH), shrubs, and vines will be measured in microplots. The microplots for measuring saplings, shrubs, and vines will have a radius of 7.3 m and microplots for measuring oak seedlings will have a 2.1-m radius. Plot sizes should be kept consistent across years and savanna areas to ensure validity of analyses.

The number of plots (of each size) sampled within each savanna area will be based on the results of these power analyses. The estimated number of plots will likely vary depending on the type of plot, the vegetation metric, and the question being answered. Placement of these plots is important to ensure independence among plots and ease of sampling. The statistical methods presented here assume that plots are independent. Using a systematic design for the tree macroplots would ensure that the savanna areas are evenly sampled and plots are independent. The plan of the monitoring study is to then locate microplots and nested frequency plots within the macroplots. If sample size estimates for variables measured in microplots or nested frequency plots are equivalent to those for variables measured in macroplots, then microplots or nested frequency plots could be placed as one per macroplot and can be assumed to be independent. However, if the number of microplots or nested frequency plots needed are greater than the number of macroplots then there are a several options: (1) arrange microplots or nested frequency plots independently of macroplots, (2) put only one microplot or nested frequency plot within each macroplot with additional plots between macroplots, or (3) put multiple microplots or nested frequency plots within a macroplot. With the first two options, the assumption of independence should be valid. For the third option, the plots might not be independent—it would depend on how far apart they could be placed. Placing them as far apart as possible will help ensure they approach independence.

The first question of interest for the savanna monitoring study concerns the health of a savanna area, which cannot be addressed with the power analysis. To assess whether a savanna area is healthy, the range of values that constitute a healthy savanna will need to be determined for each vegetation metric. A graph might be a good visual method to monitor health with time (fig. 1); for example, for metric *i* the desired values for a healthy savanna would lie between *a* and *b* and the area between *a* and *b* can be shaded. The yearly observed values can then be plotted to see if they lie within, above, or below the desired values. The plot could be created by plotting either the individual plot values or the average across all plots (as was done in fig. 1); creating both plots would give the best assessment of the health of a savanna. For example, in figure 1, the average value for the metric is above the desired value the first 2 years but then is within

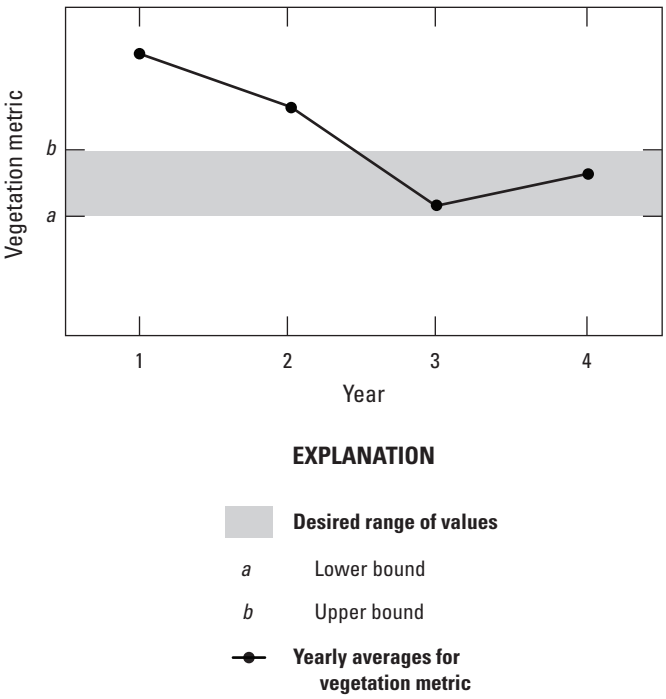

Figure 1. Example graph for monitoring the health of a savanna area.

the desired range the third and fourth years. A graph with the individual plot values could be used to indicate if most plot values are within the desired range. Even if the average value is within the desired range, having most individual plot values outside the desired range may indicate the savanna area is not as healthy as desired.

Two other questions of interest for this study concern how a savanna area changes through time and how one savanna area compares to another savanna area. Sample sizes needed for these two questions will be addressed with power analysis methods. To address the change through time for a savanna area, data from 2 or more years will need to be collected from the same plots within a savanna area. Sampling the same plots each year and using a repeated measures model will require smaller sample sizes than sampling new plots each year. To compare two savanna areas, data from two different areas within the same year should be collected.

Statistical Models

The first step in a power analysis is to define the statistical models that could be used to address the above questions. To assess changes in time for a savanna area, a generalized linear mixed model (GLMM; Stroup, 2013; Stroup and others, 2018) could be used to determine if the vegetation metric significantly differs among years for a savanna area. A GLMM consists of three components: a random component specifying

the conditional distribution of the vegetation metric, a linear predictor, and a link function. In the mixed model case, the linear predictor consists of fixed effects and random effects. The model is as follows:

$$g(\mu_{ij}) = \mu + \tau_i + \gamma_j, \quad (1)$$

where

$g(\mu_{ij})$ is the link function, which is equal to μ_{ij} for normally distributed data, $\log(\mu_{ij})$ for count data, and $\text{logit}(\mu_{ij})$ for binomial and beta-distributed data;

μ is the overall mean;

τ_i is the effect of the *i*th year; and

γ_j is the random effect of the *j*th plot.

This model is a repeated measures model with year as the repeated measure. Year is included in the model as a fixed effect and the plot, which is the subject that was sampled each year, is included as a random effect. This model will be the same for all vegetation metrics; the only difference would be the assumed distribution of each metric (see the "Vegetation Metrics" section) and the link function.

If the assumed distribution is normal or lognormal and only comparing two years, then the above model is equivalent to doing a two-sample t-test for correlated data (Graziano and Raulin, 2020; degrees of freedom=$n-1$, where n is the number of plots surveyed):

$$t = (\bar{x}_2 - \bar{x}_1) \Big/ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} - 2r\sqrt{\frac{s_1^2}{n_1}}\sqrt{\frac{s_2^2}{n_2}}}, \quad (2)$$

where

$\bar{x}_1$ is the mean for year 1;

$\bar{x}_2$ is the mean for year 2;

s_1^2 is the variance for year 1;

s_2^2 is the variance for year 2;

n_1 is the sample size (number plots) for year 1;

n_2 is the sample size for year 2; and

r is the Pearson correlation coefficient between year 1 and year 2 data.

This t-test assumes that the data are normally distributed and that the same plots were surveyed each year. This t-test can also be used for lognormal data if the means, variances, and correlation are computed on natural log-transformed data.

Assessing the difference between two or more savanna areas with respect to a vegetation metric could be done with a generalized linear model (GLM; Stroup, 2013; Stroup and others, 2018) to determine if the vegetation metric differs among savanna areas. A GLM consists of three components: a random component specifying the conditional distribution of the vegetation metric, a linear predictor, and a link function. For this model the linear predictor consists of only fixed effects.

The model is as follows:

$$g(\mu_i) = \mu + \alpha_i, \quad (3)$$

where

$g(\mu_i)$ is the link function, which is equal to μ_i for normally distributed data, $\log(\mu_i)$ for count data, and $\text{logit}(\mu_i)$ for binomial and beta-distributed data;

μ is the overall mean; and

α_i is the effect of the *i*th savanna area.

This model is a single-factor model (that is, only one factor is examined in the model); the savanna area is that single factor and is included in the model as a fixed effect. This model will be the same for all vegetation metrics; the only difference is the assumed distribution of each metric (see the "Vegetation Metrics" section) and the link function.

If the assumed distribution is normal or lognormal and the analyst is comparing two savanna areas, then the above model is equivalent to doing a two-sample t-test (Montgomery, 1997; degrees of freedom=n_1+n_2-2, where n_1=number of plots surveyed in savanna area 1 and n_2 is the number of plots surveyed in savanna area 2):

$$t = (\bar{x}_2 - \bar{x}_1) \Big/ \sqrt{\left(\frac{(n_1-1)\,s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}, \quad (4)$$

where

$\bar{x}_1$ is the mean for savanna area 1;

$\bar{x}_2$ is the mean for savanna area 2;

s_1^2 is the variance for savanna area 1;

s_2^2 is the variance for savanna area 2;

n_1 is the sample size (number plots) for savanna area 1; and

n_2 is the sample size for savanna area 2.

This t-test assumes that the plots were randomly selected from the population of interest and the data are normally distributed. This test can also be used for lognormal data if the means and variances are computed on natural log-transformed data.

In both analyses it was assumed that plots are the experimental units and, within a savanna area, plots are independent. The savanna area is the population of interest in this study. If the savanna area was the experimental unit, it would be assumed that plots within a savanna area are replicate samples and are correlated; however, given that the savanna area is the population of interest here, it should be a valid assumption that plots within a savanna area are independent if these plots are randomly or systematically placed within the savanna area.

As an alternative to using the models above, the two models could be combined into one model to simultaneously test year and savanna area effects, which would require data from two or more years from two or more savanna areas. A GLMM (Stroup, 2013; Stroup and others, 2018) could be used

to determine if the vegetation metric differs among savanna areas, among years, or among year and savanna area combinations (interaction effect). This model is a repeated measures model with year as the repeated measure. Year, savanna area, and year x savanna area would be included in the model as fixed effects, and the plot within savanna area, which is the subject that was sampled each year, is included as a random effect. This model offers an alternative to the above two models; however, it was not used in the power analyses.

Vegetation Metrics

For this power analysis, nine vegetation metrics from the savanna monitoring study were examined. Five of these metrics are to be measured in the tree macroplots, three in the microplots, and one in the nested frequency plots.

The metrics from the tree macroplots that were included are tree density, basal area, DBH, canopy cover, and fire scar ratio. Tree density (trees per hectare) and basal area (square meters per hectare) are measured for each macroplot. If a plot does not have any trees, then both metrics are zero for that plot. The DBH (centimeters) is measured for each tree within the macroplot. For analyses, DBH values for each macroplot should be averaged and these averages should be used in the statistical models. If a macroplot has no trees, then the average DBH is undefined. Based on preliminary data collected in 2020 (Pauline Drobney, U.S. Fish and Wildlife Service, written commun., 2021 [at the time of publication, data were not available from the U.S. Fish and Wildlife Service]; see the "Summary Statistics" section), these three metrics appear to have a symmetric or slightly asymmetric distribution. Therefore, in the above statistical models, a normal or lognormal distribution (Stroup, 2013) could be assumed.

Canopy cover and fire scar ratio are recorded as a percent or a proportion. Canopy cover is measured in four cardinal directions within each macroplot; these four values should be averaged to obtain an average canopy cover for each plot. Fire scar ratio is computed for each tree as the width of the fire scar divided by the circumference of the tree at the height of the fire scar. If a tree has multiple fire scars and all fire scars are at the same level (that is, have the same tree circumference), the fire scar widths could be summed and then divided by the tree circumference. Alternatively, a ratio could be computed for each fire scar and then ratios averaged for each tree or the maximum ratio used for each tree. If a tree does not have a fire scar, the fire scar ratio for that tree should be equal to zero. For each macroplot, the fire scar ratios across all trees within that plot should be averaged (that is, including trees with a fire scar ratio of zero) or the maximum fire scar ratio within that plot should be used. Because canopy cover and fire scar ratios can be recorded as proportions, a beta distribution (Stroup, 2013) can be assumed for both metrics in the above statistical models. For canopy cover, if most average proportions fall within 0.3–0.7, then it might be acceptable to assume a normal distribution, but this assumption would need to be checked. For fire scar ratio, there will likely be plots with no fire scar ratios resulting in zero values. A value of zero is not included in a beta distribution, so for fire scar ratio a zero-inflated beta distribution should be used (Ospina and Ferrari, 2010) to account for the zeros.

The vegetation metrics that are measured in the microplots are sapling counts, oak seedling counts, and shrub/vine percent cover class. The first two vegetation metrics are counts and either a Poisson or negative binomial distribution (Stroup, 2013) should be assumed in the statistical models. Data should be examined to determine which distribution is more appropriate. In a Poisson distribution the variance is equal to the mean; however, in ecological data, counts are often overdispersed, resulting in the variance being larger than expected. One way of dealing with overdispersed count data is to assume a negative binomial distribution for the counts (Stroup, 2013). To determine if a Poisson or negative binomial is most appropriate in analyses, run the analysis assuming a Poisson distribution and use a Pearson chi-square goodness-of-fit test to see if the Poisson fit is adequate. If the test shows data are overdispersed, then rerun the analysis assuming a negative binomial distribution. Saplings and oak seedlings will be counted for two size classes, which can also be combined into one; the statistical models will be the same regardless of whether the analyses were run for a size class or overall. If there are no saplings or oak seedlings in a plot, the sapling count or oak seedling count, respectively, should be zero for that plot and included in analyses.

Shrub/vine percent cover class is a categorical variable with ordered categories. There are five cover classes with the following percentage ranges: 0–10, 10–30, 30–50, 50–70, and 70–100. Cover class will be recorded for each species within the plot. Because categories cannot be averaged or summed across species, analysis for this metric will have to be done by species. If investigators prefer to analyze the overall shrub/vine cover class (regardless of species), then this overall cover class will need to be estimated in the field. Shrub/vine percent cover class is a multinomial variable; therefore, a multinomial distribution should be assumed in analyses with a cumulative logit link because the categories are ordered (Stroup, 2013).

The nested frequency plots (DeBacker and others, 2011) will be used to record the presence of herbaceous species. The presence (or absence) of a species is recorded for each plot. Frequency for each species is then computed as the number of plots in which the species was observed divided by the total number of plots. This is a binomial variable and can be analyzed assuming a binomial distribution in the above models (that is, logistic regression; Stroup, 2013).

Summary Statistics

Estimates of the expected means and variances are needed to perform a power analysis or sample size estimation. Means and variances from preliminary data collected in October 2020 at Neal Smith National Wildlife Refuge (NWR),

Jasper County, Iowa (Pauline Drobney, U.S. Fish and Wildlife Service, written commun., 2021 [at the time of publication, data were not available from the U.S. Fish and Wildlife Service]), were used for all vegetation metrics. These data were also used to help determine the distribution of each vegetation metric. However, sample sizes for some metrics were very small (<10); there were 9 tree macroplots, 9 microplots (1 per macroplot), and 36 nested frequency plots (4 per macroplot) surveyed. These small sample sizes can result in large variances and make it difficult to determine the correct distribution of the data. Therefore, two other datasets were also used to help inform what the expected means and variances for these metrics may be. One dataset was obtained from the U.S. Forest Service and consisted of data collected in 2009 from 10 plots within a savanna area in the Ha Ha Tonka State Park, Camden County, Missouri (Daniel Dey, U.S. Forest Service, written commun., 2021 [at the time of publication, data were not available from the U.S. Forest Service]). These data only included three of the vegetation metrics: tree density, basal area, and DBH. The second dataset was obtained from the U.S. Fish and Wildlife Service and consisted of data collected from savanna areas in Neal Smith NWR in 2011 (Pauline Drobney, U.S. Fish and Wildlife Service, written commun., 2021 [at the time of publication, data were not available from the U.S. Fish and Wildlife Service]). Data were collected from six savanna areas with 1–7 plots per savanna area. Because there were a small number of plots per savanna area, means and variances were calculated across all 30 plots (ignoring savanna area). Four of the vegetation metrics were included in this dataset: tree density, basal area, DBH, and canopy cover.

The summary statistics varied among datasets (table 1). The range of means and variances from these three datasets were used to help define the range of what could possibly be observed for each vegetation metric, and the power analyses were done using this defined range of means and variances.

Table 1. Summary statistics for each vegetation metric from three datasets containing data from savanna areas: 2020 Neal Smith National Wildlife Refuge and 2011 Neal Smith National Wildlife Refuge (Pauline Drobney, U.S. Fish and Wildlife Service, written commun., 2021 [at the time of publication, data were not available from the U.S. Fish and Wildlife Service]), and 2009 Ha Ha Tonka Savanna Area (Daniel Dey, U.S. Forest Service, written commun., 2021 [at the time of publication, data were not available from the U.S. Forest Service]).

[*n*, number of sample plots; SD, standard deviation; CV, coefficient of variation; m²/ha, square meter per hectare; DBH, diameter at breast height; cm, centimeter; <, less than; m, meter; >, greater than]

Metric	Subset/scale[1]	*n*	Mean	SD	Variance	CV
2020 Neal Smith National Wildlife Refuge						
Tree density	Original scale	9	255.31	230.19	52988.6	90.16
Basal area (m²/ha)	Original scale	9	18.44	10.85	117.78	58.84
DBH (cm)	Plot means, original scale	8	53.38	21.92	480.53	41.06
	Plot means, log transformed	8	3.91	0.38	0.14	9.61
Canopy cover	Plot means, percent	4	72.75	9.74	94.96	13.39
	Plot means, proportion	4	0.7275	0.0974	0.0995	13.39
Fire scar ratio	Plot means	8	0.0170	0.0329	0.0011	193.27
Sapling counts	2.5–7.6 cm DBH, all species	9	8.0	14.84	220.25	185.51
	7.6–12.7 cm DBH, all species	9	1.56	1.33	1.78	85.71
	All sizes, all species	9	9.56	14.79	218.78	154.79
Oak seedling counts	<0.9 m tall, all species	9	0.33	0.71	0.50	212.13
	>0.9 m tall, all species	9	0.78	1.72	2.94	220.62
	All sizes, all species	9	1.11	2.42	5.86	217.89
2011 Neal Smith National Wildlife Refuge						
Tree density	Original scale	30	426.43	254.96	65007.0	59.79
Basal area (m²/ha)	Original scale	30	24.56	10.62	112.73	43.22
DBH (cm)	Plot means, original scale	30	41.71	11.62	135.14	27.87
	Plot means, log transformed	30	3.70	0.26	0.07	6.93
Canopy cover	Plot means, percent	31	84.19	13.49	181.91	16.02
	Plot means, proportion	31	0.8419	0.1349	0.0182	16.02
2009 Ha Ha Tonka Savanna Area						
Tree density	Original scale	10	231.78	72.09	5196.7	31.10
Basal area (m²/ha)	Original scale	10	7.80	1.90	3.61	24.36
DBH (cm)	Plot means, original scale	10	27.39	3.36	11.31	12.28
	Plot means, log transformed	10	3.30	0.12	0.01	3.70

[1]Indicates what subset of data was used or if data were transformed.

Power Analysis

Power analysis methods were used to estimate sample sizes needed to detect significant differences between two years or between two savanna areas. The statistical models described above for each question were used in the power analyses. These models could be used for testing differences among two or more years or among two or more savanna areas; however, for the power analysis only two years or two savanna areas were compared. For each vegetation metric, sample size estimates were computed for a range of means and variances. The effect size used in the analyses varied with vegetation metric and will therefore be explained in the sections below. A significance level of α=0.05 was used in all analyses and sample sizes were determined for a power of 0.8 and 0.9 (β).

Sample size to determine if the vegetation metric differs between years for a savanna area was estimated using 2 years of data from one savanna area. It is assumed that the 2 years of data were collected from the same *n* plots each year. The number of plots (*n*) needed to detect the desired effect size with a significance level of 0.05 and a power of 0.8 or 0.9 was estimated with this analysis. Plots within a savanna area were assumed to be independent but the data from 2 years for a plot were assumed to be correlated. When computing sample sizes

needed to detect differences among years, the results depend on the correlation among years; therefore, sample sizes were estimated assuming three different levels of correlation: 0.2, 0.5, and 0.8.

To determine if the vegetation metric differs between savanna areas, the sample size was estimated using just two savanna areas. In the analyses, the number of plots in each savanna area (n) was assumed to be the same. The number of plots (n) needed to detect the desired effect size with a significance level of 0.05 and a power of 0.8 or 0.9 was estimated with this analysis. Plots within a savanna area and between savanna areas were assumed to be independent.

Sample Size Estimates for Normal and Lognormal Vegetation Metrics

Tree density, basal area, and DBH appear to be either normally distributed or lognormally distributed in the three datasets examined. The first two variables appear to be almost symmetric, so a normal distribution was assumed in analyses. Average DBH appeared to be skewed right, so a lognormal distribution was assumed for this metric. Lognormal data can be analyzed as normal if data are natural log transformed prior to analysis; therefore, for DBH, where the plot means are used in the analysis, these plot means were natural log transformed prior to computing overall means and variances.

To compute sample sizes for normal and lognormal data with the statistical models given above, the GLMPOWER procedure in SAS (SAS Institute Inc., 2018) was used. This procedure can be used to estimate sample sizes for univariate and multivariate linear models with normally distributed vegetation metrics. To use this procedure, an estimate of the mean and variance, the effect size, the significance level, the desired power, the statistical model, and the correlation among years (for comparing years) need to be specified and the procedure will then give an estimated sample size. Statistical models for both questions can be used in this procedure; the model for comparing years is a multivariate linear model and the univariate linear model is used for comparing savanna areas.

In these analyses two means were compared, either two correlated means when comparing years or two independent means when comparing savanna areas; therefore, the above statistical models are equivalent to doing two sample t-tests (as described in the "Statistical Models" section). The t-test formulas given in equations 2 and 4 were solved for n and can be used to compute an estimate of the sample size. In the power analyses it was assumed that $s_1^2=s_2^2$ and $n_1=n_2$. Then, solving the t-test equation for comparing two years (eq. 2) for n and accounting for the significance level and power level, the result was as follows:

$$n = 2s^2(t_{a/2}+t_\beta)^2(1-r)/(\bar{x}_2-\bar{x}_1)^2, \quad (5)$$

where

n is the sample size (number plots) for each year;
$\bar{x}_1$ is the mean for year 1;
$\bar{x}_2$ is the mean for year 2;
s^2 is the variance for both year 1 and year 2;
r is the Pearson correlation coefficient between year 1 and year 2 data;
$t_{a/2}$ is the t-value for the significance level (α/2); and
t_β is the t-value for the power level (β).

Solving the t-test equation for comparing two savanna areas (eq. 4) for n and accounting for both the significance level and power level gives:

$$n = 2s^2(t_{a/2}+t_\beta)^2/(\bar{x}_2-\bar{x}_1)^2, \quad (6)$$

where

n is the sample size (number plots) for each savanna area;
$\bar{x}_1$ is the mean for savanna area 1;
$\bar{x}_2$ is the mean for savanna area 2;
s^2 is the variance for both savanna area 1 and savanna area 2;
$t_{a/2}$ is the t-value for the significance level (α/2); and
t_β is the t-value for the power level (β).

The means and variances used in these analyses were based on the range of means and variances observed in the three datasets described above. Effect sizes of 10-, 20-, 30-, 40-, and 50-percent difference were used for each of these metrics. These percentages given in decimal form will be referred to as proportion difference (d); that is, in the power analyses it was assumed the second mean (the mean for year 2 or mean for the second savanna area) was equal to the first mean+(first mean x d), where d equals 0.1, 0.2, 0.3, 0.4, or 0.5. For lognormal data, which will be log transformed in the analyses, this effect size on the log scale was equivalent to the first mean+natural log (1+d).

The coefficient of variation (CV) is equal to the standard deviation (s) divided by the mean ($\bar{x}$), and therefore $s = \bar{x}_1(CV_1)$ where $\bar{x}_1$ is mean for the first year or savanna area and CV_1 is the coefficient of variation for the first year or savanna area. Then, given the definition of effect size above, $\bar{x}_2 = \bar{x}_1 + d(\bar{x}_1)$, the formula estimating n for comparing years for normally distributed variables (eq. 5) was rewritten as follows:

$$n = 2(CV_1)^2(t_{a/2}+t_\beta)^2(1-r)/d^2, \quad (7)$$

where

n is the sample size (number plots) for each year;
CV_1 is the coefficient of variation for year 1;
d is proportion difference between year 1 and year 2 means;

r is the Pearson correlation coefficient between year 1 and year 2 data;
$t_{a/2}$ is the t-value for the significance level (α/2); and
t_{β} is the t-value for the power level (β).

Therefore, for normal data, the CV for year 1 data is what matters when computing the sample size. If the CV is the same, sample size estimates will be the same regardless of the mean and standard deviation; that is, the sample size estimate for a mean of 100 and standard deviation of 10 (CV=0.1) will be the same as for a mean of 15 and standard deviation of 1.5 (CV=0.1). Therefore, power analyses were run using the means calculated from the 2020 data and several CVs based on the range of CVs from all three datasets. The formula is the same for comparing two savanna areas except (1–r) is not in the formula and all terms in the formula are defined in reference to savanna area 1 and 2 rather than year 1 and 2.

For lognormal data, $\bar{x}_2 = \bar{x}_1 + \ln(1+d)$. The formula estimating n for comparing years for lognormal variables (eq. 5) was then rewritten as follows:

$$n = 2s^2(t_{a/2}+t_{\beta})^2(1-r)/(\ln(1+d))^2, \quad (8)$$

where

n is the sample size (number plots) for each year;
s^2 is the variance for both year 1 and year 2;
d is proportion difference between year 1 and year 2 means;
r is the Pearson correlation coefficient between year 1 and year 2 data;
$t_{a/2}$ is the t-value for the significance level (α/2); and
t_{β} is the t-value for the power level (β).

Therefore, for lognormal data, the variance for year 1 data is what matters when computing the sample size. If the variance is the same, sample size estimates will be the same regardless of the mean. Therefore, power analyses were run using the means calculated from the 2020 data and several variances based on the range of variances from all three datasets. The formula is the same for comparing two savanna areas except (1–r) is not in the formula and all terms in the formula are defined in reference to savanna area 1 and 2 rather than year 1 and 2.

Sample Size Estimates for Vegetation Metrics of Other Distributions

Several of the vegetation metrics (canopy cover, sapling counts, oak seedling counts, frequency, and shrub/vine percent cover) follow distributions other than normal or lognormal, such as beta, Poisson, negative binomial, binomial, and multinomial. For these distributions, sample sizes were estimated using simulated data (Wicklin, 2013). First, data were simulated for a given distribution, sample size (n), estimated mean, variance, and effect size (and correlation when comparing years). For each sample, the appropriate model was then run using the GLIMMIX procedure in SAS (SAS Institute Inc., 2018) and whether the null hypothesis was rejected at a significance level of 0.05 was recorded. This process was then repeated for a total of 1,000 runs. The power was then equal to the proportion of runs in which the null hypothesis was rejected. This process was done for a range of sample sizes between a minimum of 2–12 (depending on distribution and test) and maximum of 120, at intervals from 1 to 4. The power from these various sample sizes was then used to estimate the sample sizes at which power was approximately equal to 0.9 and 0.8. Because every sample size between 2 and 120 was not run, the sample size with power equal to 0.9 or 0.8 had to be interpolated using a linear interpolation in some cases. For example, if the power was 0.88 with a sample size of 36 and 0.92 with a sample size of 40, the sample size where power was equal to 0.9 was estimated to be 38. This simulation process was repeated for a range of means, variances, effect sizes, and correlations (if comparing years).

Canopy cover will be measured as a percent, but it can be converted to a proportion and a beta distribution assumed. Simulations were run using a range of sample sizes from 2 to 120 and using a range of means and variances. Effect sizes of absolute values of 0.05, 0.10, 0.15, 0.20, and 0.25 were used in analyses.

Fire scar ratio will also be a proportion, but because of the presence of zeros, a zero-inflated beta should be used for analyses. In the data collected in the fall of 2020, there were few instances of fire scars, resulting in poor estimates of the means and variances. A zero-inflated beta is a more complex model than a beta distribution; given the number of model failures encountered when running the beta model for canopy cover (see "Results and Discussion" section), it is likely that more model difficulties would be encountered running the simulation with a zero-inflated beta. Therefore, sample sizes were not estimated for fire scar ratio.

Sapling counts and oak seedling counts are both count variables and either a Poisson or negative binomial distribution would be appropriate. In ecological data, counts are often over-dispersed (that is, variance is larger than expected) resulting in a poor fit if assuming Poisson. A good method of dealing with this overdispersion is to assume a negative binomial distribution instead (Stroup, 2013). Therefore, simulations were done using both distributions. Simulations were run using a range of sample sizes from a minimum of 4 (Poisson) or 8–12 (negative binomial) to a maximum of 120. Simulations often failed to converge at low sample sizes; therefore, the minimum needed to be raised sometimes (hence the range of 8–12 for the negative binomial distribution). Sample size estimation was done using a range of means and variances. First, it was assumed the variance equaled the mean and a Poisson distribution was assumed for the models. Then, a negative binomial was used with the variances equal to multiples of the mean, because in count data the variance

is a function of the mean. Effect sizes for both metrics were absolute counts with a difference of 1, 2, 3, 4, and 5 used for sapling counts and a difference of 0.5, 1.0, 1.5, 2.0, and 2.5 used for oak seedling counts.

Frequency from nested plots has a binomial distribution, with the values for each plot being binary variables (that is, present or absent). Simulations were run using a range of sample sizes from 4 to 120. Sample size estimation was done for a range of mean binomial probabilities (p) from 0.1 to 0.4 and effect sizes of absolute values of 0.1, 0.2, 0.3, 0.4, and 0.5. For these simulations, data were simulated for a given sample size (n), probability (p), effect size, and correlation (if comparing years). However, the correlation between two binary variables is constrained by the expected values of these two metrics (Wicklin, 2013); that is, not all pairwise correlations and marginal probabilities are consistent with each other. Therefore, when simulating correlated binomial data for comparing two years, not all combinations of probabilities and correlation coefficients specified can be achieved, and so simulations were only run for those combinations that were valid.

For shrub/vine percent cover class, the data are categorical and could be analyzed with a multinomial model. For the power analyses, three sets of starting probabilities for the five classes were defined. One set was similar to proportions seen for a common species in the 2020 data (that is, class 1 had a large proportion with the other categories decreasing in proportions from there). The second set had class 2 and 3 with the highest probabilities and the third set had equal probabilities for all classes. Effect sizes of absolute values of 0.05, 0.10, and 0.15 were used, with each effect size being applied by decreasing the probabilities for classes 1 and 2 by the specified amount, keeping probability for class 3 the same, and increasing probabilities for classes 4 and 5 by the specified amount. For example, if the starting probabilities are 0.15, 0.35, 0.25, 0.13, and 0.12, these probabilities were compared to 0.10, 0.30, 0.25, 0.18, and 0.17 when the effect size is 0.05.

Results and Discussion

The estimated sample sizes from the power analyses varied across vegetation metrics, and varied with mean, variance, effect size, and correlation within a vegetation metric. Simulations were run for a range of means and variances, for five effect sizes, and for three correlation levels (when comparing years) for each vegetation metric, which resulted in an estimated sample size for 80- and 90-percent power for each mean, variance, effect size, and correlation combination. These sample sizes are reported in tables 2–9 for each vegetation metric. It may be difficult to determine a single sample size for all vegetation metrics within a particular plot size that is sufficient to detect differences but not so large that sampling is difficult to complete; therefore, tables 2–9 should be used as a guide to determine sample sizes that will be large enough to detect differences for some vegetation metrics. Then for each vegetation metric, the sample size, mean, variance, and correlation can be used to determine what effect size may be detectable (see "Tree Density" section for an example).

For all vegetation metrics, estimated sample sizes are larger when comparing savanna areas where all data are uncorrelated than when comparing years that have correlated data from repeated measurements of the same plots. When comparing years, higher correlation between years will result in lower estimated sample sizes. Also, the more variable the data are (that is, higher variance), the larger the estimated sample size; all these results are as expected.

Tree Density

Means and variances for tree density differed considerably between the datasets used for summary statistics (table 1). Some data were highly variable with CVs as large as 90 percent. Because a normal distribution was assumed for this metric, the sample size will depend on the CV rather than the mean and variance; therefore, sample sizes were estimated using a mean of 255 and five CV values: 30, 45, 60, 75, and 90 percent (these correspond to standard deviations of 76.5, 114.75, 153.0, 191.25, and 229.5 when the mean is 255).

The number of macroplots needed to detect a specified difference in tree density between savanna units ranged from 7 to more than 1,700; and the number needed to detect a difference between years ranged from 4 to more than 1,300 (table 2). If, for example, the goal is to detect a 20-percent difference between yearly means, with 90-percent power, and the data have low variation (CV=about 30 percent), then the number of samples needed ranges from 12 to 40 plots. But if the data are highly variable (CV=about 90 percent), this would change to 88–343 plots. As stated in the previous section, the tables can be used as a guide to estimate sample sizes that will be large enough to detect some differences but not so large that sampling is difficult to complete; and then the tables can be used to determine what effect size may be detectable with that sample size. For example, assume there are 30 tree macroplots and it is expected that the tree density for a given savanna area will be moderately variable (CV=about 60 percent) and that the 2 years of data will be moderately correlated (r=0.5). Then, with 30 plots, CV=60 percent, and r=0.5, it will be possible to detect approximately a 40-percent difference in tree density with 90-percent power and a 30-percent difference with 80-percent power.

Basal Area

Basal area means and variances were somewhat similar for the two Neal Smith NWR datasets but were considerably lower at the Ha Ha Tonka savanna area (table 1). Data at all three areas were fairly symmetric, so a normal distribution was assumed in analyses. Therefore, sample sizes will depend on just the CV and were estimated using a mean of 18.4 and five

Table 2. Sample size estimates for detecting a difference in tree density (all species combined) between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; PDiff, percent difference; CV, coefficient of variation, in percent; n/a, not applicable]

r	PDiff	Sample size[1] at 90-percent power					Sample size[1] at 80-percent power				
		CV=30	CV=45	CV=60	CV=75	CV=90	CV=30	CV=45	CV=60	CV=75	CV=90
Comparing savanna areas											
n/a	10	191	427	758	1,184	1,704	143	319	567	884	1,273
	20	49	108	191	297	427	37	81	143	222	319
	30	23	49	86	133	191	17	37	64	100	143
	40	13	28	49	75	108	10	21	37	57	81
	50	9	19	32	49	70	7	14	24	37	52
Comparing years											
0.2	10	154	343	608	948	1,364	115	257	455	709	1,020
	20	40	88	154	239	343	31	66	115	179	257
	30	19	40	70	108	154	15	31	53	81	115
	40	12	24	40	62	88	10	18	31	47	66
	50	9	16	27	40	57	7	13	21	31	43
0.5	10	97	215	381	593	854	73	161	285	444	638
	20	26	56	97	150	215	20	42	73	113	161
	30	13	26	44	68	97	10	20	34	52	73
	40	9	16	26	39	56	7	12	20	30	42
	50	7	11	18	26	37	6	9	14	20	28
0.8	10	40	88	154	239	343	31	66	115	179	257
	20	12	24	40	62	88	10	18	31	47	66
	30	7	12	19	29	40	6	10	15	22	31
	40	5	8	12	17	24	5	7	10	14	18
	50	4	6	9	12	16	4	5	7	10	13

[1]The sample sizes are the number of tree macroplots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

CV values: 24, 33, 42, 51, and 60 percent (these correspond to standard deviations of 4.42, 6.07, 7.73, 9.38, and 11.04 when the mean is 18.4).

Basal area was a little less variable than tree density, resulting in lower sample size estimates (table 3); however, when data are normally distributed, sample size only depends on CV. Thus, the results in table 3 are the same as in table 2, where CV is the same (for example, CV=60 percent). The only difference between the two tables is that the levels of variation used in the power analyses were smaller for basal area than for tree density.

Diameter at Breast Height (DBH)

In all three datasets examined for summary statistics (table 1), mean DBH was skewed to the right (that is, distribution has a long tail to the right); therefore, a lognormal distribution was assumed in analyses for DBH. As with basal area and tree density, the mean and variance were lower at Ha Ha Tonka than at Neal Smith NWR. The range of values from these three datasets was used to define the variance used in the power analyses. As discussed in the "Sample Size Estimates for Normal and Lognormal Vegetation Metrics" section, in the lognormal case, the sample size will just depend on the variance; therefore, sample sizes were estimated using a mean of 3.91 and five variance values: 0.014, 0.036, 0.068, 0.109, and 0.160 (standard deviations of 0.12, 0.19, 0.26, 0.33, and 0.40 respectively) (all these values are on a log scale).

Sample size requirements for mean DBH are smaller than for the previous two metrics (table 4), which is due to less variable data used in the analyses; CVs for the DBH means and variances used ranged from 3 to 10 percent. The data are less variable because they are plot means rather than individual values. A 30-percent difference between means could be detected with <50 plots at all variance levels.

Table 3. Sample size estimates for detecting a difference in basal area (all species combined) between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; PDiff, percent difference; CV, coefficient of variation, in percent; n/a, not applicable]

r	PDiff	Sample size[1] at 90-percent power					Sample size[1] at 80-percent power				
		CV=24	CV=33	CV=42	CV=51	CV=60	CV=24	CV=33	CV=42	CV=51	CV=60
Comparing savanna areas											
n/a	10	123	230	372	548	758	92	172	278	410	567
	20	32	59	94	138	191	24	44	71	104	143
	30	15	27	43	62	86	12	21	32	47	64
	40	9	16	25	36	49	7	12	19	27	37
	50	6	11	16	23	32	5	8	13	18	24
Comparing years											
0.2	10	99	186	299	440	608	75	139	224	329	455
	20	27	48	77	112	154	21	37	58	84	115
	30	13	23	35	51	70	11	18	27	39	53
	40	9	14	21	30	40	7	11	16	23	31
	50	7	10	14	20	27	6	8	11	16	21
0.5	10	63	117	188	276	381	48	88	141	207	285
	20	18	31	49	71	97	14	24	37	53	73
	30	9	15	23	33	44	8	12	18	25	34
	40	7	10	14	20	26	6	8	11	15	20
	50	5	7	10	14	18	5	6	8	11	14
0.8	10	27	48	77	112	154	21	37	58	84	115
	20	9	14	21	30	40	7	11	16	23	31
	30	5	8	11	15	19	5	6	9	12	15
	40	4	6	7	9	12	4	5	6	8	10
	50	4	5	6	7	9	3	4	5	6	7

[1]The sample sizes are the number of tree macroplots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

Canopy Cover

Mean canopy cover was >70 percent in both datasets used for summary statistics that contained this metric (table 1), and plot values were all >50 percent. With all values in the upper end of the possible range, the distribution of values was skewed in both datasets; therefore, percent canopy cover was converted to a proportion and assumed to follow a beta distribution. Simulations to estimate sample sizes were run for three mean levels and three variance levels for each mean. The means used were 0.72, 0.78, and 0.84 and the variance levels used were 0.0081, 0.0169, and 0.0289.

Sample sizes needed to detect a specified difference increased with variance but decreased with an increase in mean (table 5). Because power analyses were run with a maximum sample size of 120, the exact sample size with 80- or 90-percent power could not be estimated if the power at a sample size of 120 was less than 0.8 or 0.9, respectively; therefore, these sample size estimates were recorded as >120. When estimated sample sizes were low, there was little or no difference in the estimated sample size for different means when all other factors were held constant. Sample sizes of 16–54 would be required to detect a difference in means of 0.1 between two savanna areas with 90-percent power. Sample sizes of 14–46 with low correlation, but only 6–15 with high correlation, would be required to detect a difference of the same value between two years.

Sapling Counts

Saplings were counted for two size categories in the 2020 survey at Neal Smith NWR (table 1). Summary statistics for each size category and for the two size categories combined were examined and used to define the range of means used in the simulations to estimate sample sizes. For count data, variances are a function of the mean (for example, in a Poisson distribution the variance is equal to the mean). Therefore, instead of using a range of variances

Table 4. Sample size estimates for detecting a difference in mean DBH (all species combined) between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; PDiff, percent difference; Var, variance; n/a, not applicable]

r	PDiff	Sample size[1] at 90-percent power					Sample size[1] at 80-percent power				
		Var= 0.014	Var= 0.036	Var= 0.068	Var= 0.109	Var= 0.160	Var= 0.014	Var= 0.036	Var= 0.068	Var= 0.109	Var= 0.160
Comparing savanna areas											
n/a	10	35	85	158	253	372	26	64	118	190	278
	20	11	24	44	70	103	8	19	33	53	77
	30	6	13	22	35	50	5	10	17	26	38
	40	4	8	14	22	31	4	7	11	17	24
	50	4	6	10	15	22	3	5	8	12	17
Comparing years											
0.2	10	29	69	128	204	299	22	52	96	153	224
	20	10	21	37	58	83	8	16	28	44	63
	30	6	11	19	29	42	5	9	15	22	32
	40	5	8	13	19	26	4	7	10	15	20
	50	4	6	10	14	19	4	5	8	11	15
0.5	10	19	44	81	128	188	15	34	61	97	141
	20	7	14	24	37	53	6	11	18	28	40
	30	5	8	13	19	27	4	7	10	15	21
	40	4	6	9	13	17	4	5	7	10	14
	50	4	5	7	10	13	3	4	6	8	10
0.8	10	9	19	34	53	76	8	15	26	40	58
	20	5	7	11	16	23	4	6	9	13	18
	30	4	5	7	9	12	3	4	6	8	10
	40	3	4	5	7	9	3	4	5	6	7
	50	3	4	5	6	7	3	3	4	5	6

[1]The sample sizes are the number of tree macroplots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

in the simulations, the variance was a multiple of the mean; four levels of a multiplier were used in the simulations. The mean levels used in the simulations were 2, 6, and 10. The four multipliers for the variance were 1, 5, 11, and 18. When the multiplier was 1, a Poisson distribution was assumed and for the other three multipliers a negative binomial distribution was assumed.

Sample sizes varied considerably among variance levels and means (table 6). Sample size increased with the mean and variance. Sample sizes are lowest if the variance is equal to the mean and a Poisson distribution can be assumed. As stated in the "Sample Size Estimates for Vegetation Metrics of Other Distributions" section, in ecological data, counts are often overdispersed, so a Poisson distribution is often not valid; therefore, the results using a negative binomial distribution are probably more useful. When assuming a negative binomial distribution, a difference of 3 with low variance could be detected, with 90-percent power, using a sample size of 12–40 microplots if the mean is low; however, as the mean or the variance increases the sample size needed will also increase substantially.

The power analyses were run with a maximum sample size of 120, so the exact sample size with 80- or 90-percent power could not be estimated if the power at a sample size of 120 was less than 0.8 or 0.9, respectively; therefore, these sample size estimates were recorded as >120. When running simulations, an attempt was made to compute power for sample sizes down to a minimum of 8 or 12. If the power at these low sample sizes was >0.8 or 0.9, the exact sample for 80- or 90-percent power, respectively, could not be determined and the sample size was recorded as less than or equal to (≤) 8 or ≤12. However, these analyses with low sample sizes were difficult to run and many models failed to converge. Therefore, any estimated sample size <20 may not be accurate. Poisson regression and negative binomial regression (that is, GLM or GLMM with a Poisson or negative binomial distribution)

Table 5. Sample size estimates for detecting a difference in mean proportion canopy cover between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; Var, variance; n/a, not applicable; >, greater than]

r	Difference	Mean	Sample size[1] at 90-percent power			Sample size[1] at 80-percent power		
			Var=0.0081	Var=0.0169	Var=0.0289	Var=0.0081	Var=0.0169	Var=0.0289
Comparing savanna areas								
n/a	0.05	0.72	64	>120	>120	47	96	>120
		0.78	60	114	>120	48	84	120
		0.84	57	89	102	41	65	78
	0.10	0.72	19	36	54	13	25	40
		0.78	17	32	46	13	25	36
		0.84	16	26	31	12	19	24
	0.15	0.72	9	16	26	7	12	20
		0.78	9	15	23	7	11	18
		0.84	8	14	18	6	10	14
	0.20	0.72	5	10	15	4	8	12
		0.78	6	10	14	4	7	11
		0.84	6	9	12	4	7	10
	0.25	0.72	4	7	11	3	5	8
		0.78	4	7	11	4	5	8
		0.84	4	7	9	3	5	7
Comparing years								
0.2	0.05	0.72	54	108	>120	42	78	120
		0.78	53	97	>120	39	76	107
		0.84	47	78	98	36	61	73
	0.10	0.72	16	28	46	11	23	34
		0.78	14	26	41	12	21	31
		0.84	14	24	31	11	18	24
	0.15	0.72	8	14	23	6	10	17
		0.78	8	14	21	6	11	15
		0.84	8	12	17	6	10	13
	0.20	0.72	6	9	13	5	7	11
		0.78	5	9	13	5	7	10
		0.84	6	9	12	5	7	9
	0.25	0.72	4	7	10	4	5	8
		0.78	4	7	9	4	6	8
		0.84	5	7	9	4	6	7

Table 5. Sample size estimates for detecting a difference in mean proportion canopy cover between two years for a savanna area at three correlation levels, or between two savanna areas.—Continued

[*r*, correlation level; Var, variance; n/a, not applicable; >, greater than]

r	Difference	Mean	Sample size[1] at 90-percent power			Sample size[1] at 80-percent power		
			Var=0.0081	Var=0.0169	Var=0.0289	Var=0.0081	Var=0.0169	Var=0.0289
Comparing years—Continued								
		0.72	35	71	114	27	52	91
	0.05	0.78	35	66	100	28	51	79
		0.84	34	58	73	25	43	61
		0.72	10	19	31	8	15	24
	0.10	0.78	10	19	29	8	15	22
		0.84	11	18	25	8	14	18
		0.72	6	10	15	5	8	11
0.5	0.15	0.78	6	10	14	5	8	12
		0.84	6	10	14	5	8	11
		0.72	5	7	10	4	6	8
	0.20	0.78	5	7	10	4	6	8
		0.84	5	7	9	4	6	8
		0.72	4	5	7	3	4	6
	0.25	0.78	4	6	7	4	5	6
		0.84	4	6	8	4	5	6
		0.72	16	30	49	12	23	39
	0.05	0.78	16	28	47	12	23	35
		0.84	16	30	42	13	22	33
		0.72	6	10	15	5	8	11
	0.10	0.78	6	10	15	5	8	12
		0.84	6	10	13	5	8	11
		0.72	4	6	8	4	5	6
0.8	0.15	0.78	4	6	8	4	5	7
		0.84	4	7	9	4	5	7
		0.72	3	4	6	3	4	5
	0.20	0.78	4	5	6	3	4	5
		0.84	4	5	6	4	4	5
		0.72	3	4	5	3	3	4
	0.25	0.78	3	4	5	3	4	4
		0.84	4	4	5	3	4	5

[1]The sample sizes are the number of tree macroplots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

are large sample procedures; therefore, sample sizes <20 (and perhaps even <30) may result in models not converging, computational issues, and unreliable results.

Oak Seedling Counts

Like sapling counts, oak seedling counts were counted for two size categories in the 2020 survey at Neal Smith NWR (table 1). Summary statistics for each size category and for the two size categories combined were examined and used to define the range of means used in the simulations to estimate sample sizes. Similar to results in the "Sapling Counts" section, the variances in these simulations were a multiple of the mean; four levels of a multiplier were used. The mean levels used in the simulations were 0.3, 0.9, and 1.5. The four multipliers for the variance were 1, 2, 4, and 6. When the multiplier was 1, a Poisson distribution was assumed and for the other three multipliers a negative binomial distribution was assumed.

Similar to sapling counts, sample sizes for oak seedling counts increased with increasing mean and variance (table 7). Sample sizes are lowest if the analyst is able to assume a Poisson distribution. However, like sapling counts, oak seedling counts will likely be overdispersed, so it would be best to use sample sizes estimated assuming a negative binomial distribution. Therefore, assuming a negative binomial distribution, a difference of 2 could be detected, with 90-percent power, using sample sizes of 15–75 microplots when comparing savanna areas or 12–63 microplots when comparing years.

For these simulations, a minimum sample size of 12 was used and resulted in more challenges than when running the analyses for sapling counts. Many models failed to converge, especially at the lower sample sizes; therefore, any estimated sample size <20 may not be accurate. Poisson and negative binomial regressions are large sample procedures; therefore, sample sizes <20 (and perhaps even <30) may result in models not converging or otherwise unreliable results. Like sapling counts, sample size estimates of >120 or ≤12 were used where more exact sample sizes could not be interpolated.

Frequency

Frequency from nested plots was treated as a binomial variable in sample size simulations. When using the nested plot method, the optimal subplot size to use is that which gives a frequency between 0.2 and 0.5 (DeBacker and others, 2011); therefore, probabilities (p) equal to 0.1, 0.2, 0.3, and 0.4 were used for analyses. Even though p=0.5 is part of the optimal range, it was not considered here because of the range of differences tested; the largest difference used was 0.5, which would have given the second savanna area or year a p of 1.0 and could have caused issues when running the simulations.

Sample size requirements were usually 20 or more, but they were often much larger than 20, especially for detecting smaller differences (table 8), which is expected given that logistic regression (that is, GLM or GLMM with a binomial distribution) is a large sample procedure. Generally, a sample should consist of at least 10 presences and 10 absences for this test, giving a minimum sample size of at least 20 if p=0.5. If p=0.3, the minimum sample size should be at least 33 to ensure at least 10 presences and 10 absences. When simulating binomial data, not every pairwise correlation can be achieved for a given pair of probabilities; therefore, some combinations of probability and correlation used in this simulation design are not possible and no sample size results are provided for these combinations (table 8). Because power analyses were run with a maximum sample size of 120, the exact sample size with 80- or 90-percent power could not be estimated if the power at a sample size of 120 was <0.8 or 0.9, respectively; therefore, these sample size estimates were recorded as >120.

Shrub/Vine Percent Cover Class

Shrub and vine percent cover class was treated as a multinomial variable in the power analyses. Multinomial data can be summarized to the proportion of observations that occurred in each class. These proportions need to sum to one across all five classes; therefore, it is not possible to increase each starting probability by some amount as was done for the means, counts, or frequencies of other metrics. Also, specifying the starting conditions is not as simple as specifying a mean and variance, so, for this metric, three starting conditions were specified. The first starting condition is based on data for a common species from the 2020 Neal Smith NWR dataset; these starting probabilities were p_1=0.78, p_2=0.16, p_3=0.04, p_4=0.01, and p_5=0.01. The second starting condition had higher probabilities for the second and third classes: p_1=0.15, p_2=0.35, p_3=0.25, p_4=0.13, and p_5=0.12. The third starting condition had equal probabilities of 0.2 for all classes. Effect sizes were applied by decreasing p_1 and p_2 by the specified amount, keeping p_3 the same, and increasing p_4 and p_5 by the specified amount.

Only one of the starting conditions was based on actual data. There are endless possibilities for these starting conditions and without much data on which to base these conditions, estimates of possible starting proportions for savanna areas monitored using the protocols to be developed are unclear. Also, it was unclear what type of effect sizes are possible. The sample size estimates will depend on these starting conditions, the effect size, and how these effect sizes are applied; therefore, these simulations were only run to test for differences between two savanna areas (that is, using uncorrelated data). The samples size estimates for these tests indicate that the results depend on the starting condition (table 9), and sample size estimates decrease as effect size increases (as expected). If these same analyses were run using correlated data to test for differences among years, the sample

Table 6. Sample size estimates for detecting a difference in sapling counts between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; Var, variance; n/a, not applicable; >, greater than; ≤, less than or equal to]

r	Difference	Mean	Sample size[1] at 90-percent power				Sample size[1] at 80-percent power			
			Var=1x	Var=5x	Var=11x	Var=18x	Var=1x	Var=5x	Var=11x	Var=18x
Comparing savanna areas										
n/a	1	2	54	>120	>120	>120	42	>120	>120	>120
		6	>120	>120	>120	>120	104	>120	>120	>120
		10	>120	>120	>120	>120	>120	>120	>120	>120
	2	2	17	79	>120	>120	13	58	119	>120
		6	38	>120	>120	>120	28	>120	>120	>120
		10	57	>120	>120	>120	45	>120	>120	>120
	3	2	10	40	87	>120	8	30	62	103
		6	18	87	>120	>120	14	63	>120	>120
		10	28	>120	>120	>120	21	98	>120	>120
	4	2	7	26	54	89	6	19	36	60
		6	12	52	114	>120	9	41	80	>120
		10	17	79	>120	>120	13	60	>120	>120
	5	2	5	19	36	63	≤4	14	28	45
		6	9	36	79	120	7	27	57	96
		10	12	55	115	>120	9	40	90	>120
Comparing years										
0.2	1	2	51	>120	>120	>120	38	>120	>120	>120
		6	>120	>120	>120	>120	95	>120	>120	>120
		10	>120	>120	>120	>120	>120	>120	>120	>120
	2	2	17	65	>120	>120	14	47	95	>120
		6	37	>120	>120	>120	29	113	>120	>120
		10	54	>120	>120	>120	43	>120	>120	>120
	3	2	10	33	67	106	8	24	49	76
		6	18	68	>120	>120	14	53	114	>120
		10	27	111	>120	>120	21	85	>120	>120
	4	2	7	22	43	68	6	16	33	45
		6	12	44	95	>120	10	32	70	112
		10	16	66	>120	>120	13	47	108	>120
	5	2	6	16	31	47	5	12	25	34
		6	9	31	64	100	7	22	49	75
		10	12	43	92	>120	10	33	71	112

Table 6. Sample size estimates for detecting a difference in sapling counts between two years for a savanna area at three correlation levels, or between two savanna areas.—Continued

[*r*, correlation level; Var, variance; n/a, not applicable; >, greater than; ≤, less than or equal to]

r	Difference	Mean	Sample size[1] at 90-percent power				Sample size[1] at 80-percent power			
			Var=1x	Var=5x	Var=11x	Var=18x	Var=1x	Var=5x	Var=11x	Var=18x
Comparing years—Continued										
0.5	1	2	45	118	>120	>120	35	88	>120	>120
		6	109	>120	>120	>120	88	>120	>120	>120
		10	>120	>120	>120	>120	>120	>120	>120	>120
	2	2	15	37	71	114	13	30	57	88
		6	30	84	>120	>120	25	65	>120	>120
		10	48	>120	>120	>120	39	102	>120	>120
	3	2	9	20	39	60	8	16	28	45
		6	16	42	79	>120	14	32	63	98
		10	24	60	>120	>120	19	50	97	>120
	4	2	6	15	27	40	6	11	20	30
		6	10	28	50	76	9	20	39	60
		10	14	40	78	115	12	31	57	92
	5	2	6	11	20	29	5	9	15	20
		6	8	19	37	55	7	15	29	42
		10	11	28	53	83	9	20	40	60
0.8	1	2	36	59	90	>120	30	45	67	99
		6	85	>120	>120	>120	72	104	>120	>120
		10	>120	>120	>120	>120	118	>120	>120	>120
	2	2	13	20	31	43	12	16	24	32
		6	25	42	68	92	22	31	52	68
		10	38	65	111	>120	33	47	85	113
	3	2	8	12	18	25	8	10	14	19
		6	13	23	35	45	12	17	27	35
		10	19	32	50	74	17	25	41	55
	4	2	6	9	13	18	6	≤8	≤12	14
		6	9	15	23	29	8	≤12	16	23
		10	12	19	32	44	11	16	25	32
	5	2	5	≤8	≤12	15	5	≤8	≤12	≤12
		6	7	≤12	16	22	6	≤12	13	17
		10	9	15	22	31	8	≤12	17	23

[1]The sample sizes are the number of microplots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

Table 7. Sample size estimates for detecting a difference in oak seedling counts between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; Var, variance; n/a, not applicable; >, greater than; ≤, less than or equal to]

r	Difference	Mean	Sample size[1] at 90-percent power				Sample size[1] at 80-percent power			
			Var=1x	Var=2x	Var=4x	Var=6x	Var=1x	Var=2x	Var=4x	Var=6x
Comparing savanna areas										
n/a	0.5	0.3	46	85	>120	>120	36	65	>120	>120
		0.9	93	>120	>120	>120	76	>120	>120	>120
		1.5	>120	>120	>120	>120	107	>120	>120	>120
	1.0	0.3	17	32	55	86	14	23	40	60
		0.9	30	57	114	>120	24	44	85	>120
		1.5	42	79	>120	>120	33	65	>120	>120
	1.5	0.3	12	19	33	47	10	15	26	34
		0.9	16	31	61	88	12	22	43	62
		1.5	21	43	80	>120	17	33	59	93
	2.0	0.3	9	15	24	35	8	12	19	26
		0.9	11	20	37	56	9	15	27	40
		1.5	14	27	52	75	11	20	38	58
	2.5	0.3	9	13	22	27	7	11	16	22
		0.9	8	15	26	39	7	11	19	30
		1.5	10	19	35	52	9	14	27	37
Comparing years										
0.2	0.5	0.3	45	75	>120	>120	34	57	106	>120
		0.9	90	>120	>120	>120	69	117	>120	>120
		1.5	>120	>120	>120	>120	102	>120	>120	>120
	1.0	0.3	17	27	48	65	15	21	33	49
		0.9	30	46	87	>120	23	37	73	99
		1.5	42	69	>120	>120	31	53	101	>120
	1.5	0.3	12	18	28	38	10	14	22	30
		0.9	16	26	48	69	13	20	37	52
		1.5	22	35	68	99	17	26	49	76
	2.0	0.3	10	14	23	31	8	≤12	18	23
		0.9	11	18	31	45	10	14	23	32
		1.5	13	23	42	63	12	17	31	47
	2.5	0.3	9	13	20	26	7	≤12	16	20
		0.9	9	13	23	33	7	≤12	17	23
		1.5	10	16	30	42	9	13	20	33

Table 7. Sample size estimates for detecting a difference in oak seedling counts between two years for a savanna area at three correlation levels, or between two savanna areas.—Continued

[*r*, correlation level; Var, variance; n/a, not applicable; >, greater than; ≤, less than or equal to]

r	Difference	Mean	Sample size[1] at 90-percent power				Sample size[1] at 80-percent power			
			Var=1x	Var=2x	Var=4x	Var=6x	Var=1x	Var=2x	Var=4x	Var=6x
Comparing years—Continued										
		0.3	40	52	91	>120	32	42	67	94
	0.5	0.9	80	106	>120	>120	63	76	>120	>120
		1.5	>120	>120	>120	>120	97	110	>120	>120
		0.3	16	22	35	47	14	18	27	37
	1.0	0.9	25	34	57	82	21	27	43	59
		1.5	36	46	79	107	28	35	59	83
		0.3	11	16	24	32	10	12	19	25
0.5	1.5	0.9	15	19	30	43	12	15	25	32
		1.5	19	25	41	56	16	19	31	43
		0.3	9	13	20	26	8	≤12	15	20
	2.0	0.9	10	14	20	28	9	≤12	16	22
		1.5	13	18	26	37	11	14	21	29
		0.3	9	≤12	18	24	7	≤12	14	18
	2.5	0.9	8	≤12	16	21	7	≤12	≤12	17
		1.5	9	≤12	19	28	8	≤12	15	20
		0.3	35	38	52	69	29	30	42	51
	0.5	0.9	66	69	98	>120	56	52	72	92
		1.5	98	98	>120	>120	83	74	104	>120
		0.3	15	17	25	31	13	15	19	26
	1.0	0.9	22	26	33	41	19	20	25	32
		1.5	29	34	44	54	25	28	34	39
		0.3	11	13	20	26	10	≤12	16	20
0.8	1.5	0.9	12	15	19	23	11	≤12	15	19
		1.5	16	19	24	30	14	16	19	22
		0.3	9	≤12	18	22	8	≤12	14	17
	2.0	0.9	9	≤12	14	16	8	≤12	≤12	13
		1.5	11	13	16	20	10	≤12	13	16
		0.3	9	≤12	18	23	7	≤12	13	16
	2.5	0.9	7	≤12	≤12	14	7	≤12	≤12	≤12
		1.5	8	≤12	≤12	15	8	≤12	≤12	≤12

[1]The sample sizes are the number of micoplots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

Table 8. Sample size estimates for detecting a difference in herbaceous frequency between two years for a savanna area at three correlation levels, or between two savanna areas.

[*r*, correlation level; p, probability; n/a, not applicable; >, greater than; --, no data]

r	Difference	Sample size[1] at 90-percent power				Sample size[1] at 80-percent power			
		p=0.1	*p*=0.2	*p*=0.3	*p*=0.4	*p*=0.1	*p*=0.2	*p*=0.3	*p*=0.4
Comparing savanna areas									
n/a	0.1	>120	>120	>120	>120	>120	>120	>120	>120
	0.2	80	111	120	>120	67	86	94	96
	0.3	44	52	58	56	34	40	40	46
	0.4	30	30	31	30	25	25	23	23
	0.5	24	21	20	26	21	16	16	20
Comparing years									
0.2	0.1	>120	>120	>120	>120	>120	>120	>120	>120
	0.2	71	98	104	115	56	75	85	86
	0.3	42	47	53	54	35	37	41	40
	0.4	37	31	32	29	27	26	26	25
	0.5	28	23	24	28	25	19	18	25
0.5	0.1	104	>120	>120	>120	83	>120	>120	>120
	0.2	32	61	70	77	26	48	59	58
	0.3	--	28	53	51	--	20	42	42
	0.4	--	--	--	--	--	--	--	--
	0.5	--	--	--	--	--	--	--	--
0.8	0.1	--	--	58	>120	--	--	41	92
	0.2	--	--	--	--	--	--	--	--
	0.3	--	--	--	--	--	--	--	--
	0.4	--	--	--	--	--	--	--	--
	0.5	--	--	--	--	--	--	--	--

[1]The sample sizes are the number of nested frequency plots needed within each of two savanna areas when comparing savanna areas or needed within a savanna area and surveyed for 2 years when comparing years.

Table 9. Sample size estimates for detecting a difference in shrub and vine percent cover class proportions between two savanna areas.

[*r*, correlation level; n/a, not applicable; >, greater than. Starting set of probabilities for Run 1: p_1=0.78, p_2=0.16, p_3=0.04, p_4=0.01, p_5=0.01; for Run 2: p_1=0.15, p_2=0.35, p_3=0.25, p_4=0.13, p_5=0.12; and for Run 3: p_1=0.2, p_2=0.2, p_3=0.2, p_4=0.2, and p_5=0.2]

r	Effect size	Sample size[1] at 90-percent power			Sample size[1] at 80-percent power		
		Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
Comparing savanna areas							
n/a	0.05	>120	>120	>120	>120	>120	>120
	0.10	>120	89	117	>120	67	87
	0.15	91	41	49	68	29	38

[1]The sample sizes are the number of microplots needed within each of two savanna areas.

size estimates should be lower than these given estimates and continue to decrease as correlation increases. Sample size estimates of >120 were used where a more exact sample size could not be interpolated.

Summary

When designing data collection protocols for a new research project, it is important to have a large enough sample size to detect a desired effect, but not so large as to waste time collecting more data than are needed. Power analysis methods can be used to estimate this sample size. Estimated sample sizes depend on many factors including effect size, mean and variance of data, significance level, statistical power, experimental design, distribution of data, statistical model, and correlation among repeated measures. The difference in sample size estimates among vegetation metrics was indicated in the results discussed within this report, with estimated sample sizes differing for each vegetation metric and ranging by three orders of magnitude. Within each vegetation metric, the estimated sample sizes varied with means, variances, effect size, and correlation between years. It will be difficult to find a single sample size that is sufficient to detect differences in all vegetation metrics but not so large that sampling is difficult to complete; however, the tables from this report can be used to help inform what level of difference may be detectable for a given sample size, mean, variance, and correlation. For normal and lognormal distributions, sample sizes estimates were computed using GLMPOWER procedure in SAS; therefore, for a given mean, variance, effect size, significance level, and power, the sample size estimate will always be the same for every computation. For beta, Poisson, negative binomial, binomial, and multinomial distribution, the sample size estimates were based on simulated data (1,000 runs). Therefore, because simulated data were used, the sample size estimates could vary slightly every time the simulation is run.

Generalized linear models and generalized linear mixed models assuming beta, Poisson, negative binomial, binomial, or multinomial distributions are generally considered large sample size procedures; therefore, for these distributions, minimum sample sizes of 20 (or maybe even 30) should be considered. In the power analyses, an attempt was made to estimate power with lower sample sizes, but for all these distributions a percentage of the runs failed to converge. Most convergence problems occurred with generalized linear mixed models (models comparing years) and occurred mainly when assuming a negative binomial distribution or a beta distribution. For binomial, multinomial, and Poisson distribution, less than or equal to 1.2 percent of the model runs within a set of 1,000 runs resulted in nonconvergence. For beta and negative binomial distributions, as much as 9.3 percent of the models within a set of 1,000 runs failed to converge in some cases. When analyzing data collected on savanna areas, nonconvergence could be an issue if using too small a sample size; however, because small sample sizes (less than 10) were used in these simulations, more nonconvergence issues were probably encountered in these analyses.

References Cited

DeBacker, M.D., Heywood, J.S., and Morrison, L.W., 2011, Optimized frequency measures for monitoring trends in tallgrass prairie: Rangeland Ecology and Management, v. 64, no. 3, p. 301–308, accessed July 30, 2021, at https://doi.org/10.2111/REM-D-09-00179.1.

Graziano, A.M., and Raulin, M.L., 2020, Research method—A process of inquiry (9th ed.): accessed October 1, 2021, at https://graziano-raulin.com/default.htm.

Montgomery, D.C., 1997, Design and analysis of experiments (4th ed.): New York, John Wiley and Sons, 704 p.

Ospina, R., and Ferrari, S.L.P., 2010, Inflated beta distributions: Statistical Papers (Berlin, Germany), v. 51, no. 1, p. 111–126. [Also available at https://doi.org/10.1007/s00362-008-0125-4.]

SAS Institute Inc, 2018, SAS/STAT® 15.1 user's guide: Cary, N.C., SAS Institute Inc., accessed on July 30, 2021, at https://go.documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/titlepage.htm.

Steidl, R.J., Hayes, J.P., and Schauber, E., 1997, Statistical power analysis in wildlife research: The Journal of Wildlife Management, v. 61, no. 2, p. 270–279, accessed July 30, 2021, at https://doi.org/10.2307/3802582.

Stroup, W.W., 2013, Generalized linear mixed models—Modern concepts, methods and applications: Boca Raton, Florida, CRC Press, 529 p.

Stroup, W.W., Milliken, G.A., Claassen, E.A., and Wolfinger, R.D., 2018, SAS® for mixed models—Introduction and basic application: Cary, N.C., SAS Institute Inc., 594 p.

Wicklin, R., 2013, Simulating data with SAS®: Cary, N.C., SAS Institute Inc., 354 p.

Appendix 1. SAS Programs for Running Power Analyses

```
*************************************************************************************************
*                                                                                               *
*     ***** Power analysis - Compare years macros.sas                                           *
*                                                                                               *
*        *** Date first written:  27 April 2021                                                 *
*            Date last modified:  25 September 2021                                             *
*            Written by:  Deb Buhl                                                              *
*                                                                                               *
*  This program contains macros needed for running power analyses on tests to compare 2 years  *
*  for a savanna area. There are macros for running proc glmpower, macros for simulating       *
*  data from various distributions (beta, Poisson, negative binomial, binomial), and macros    *
*  for running a GLMM model using proc glimmix to compare years within a savanna area.         *
*                                                                                               *
*************************************************************************************************;


*************************************************************************************************;
*** Proc glmpower macro - normal data.
        This macro estimates the sample sizes needed to detect a 10, 20, 30, 40, and 50 percent
        difference between yearly means at a significance level of 0.05 and with 80 and 90
        percent power. In this macro, the model used for comparing years is a repeated measures
        model and assumes data are normally distributed. Sample sizes are estimated for a given
        mean and level of variation, and for three correlation levels (0.2, 0.5, and 0.8)
        between years. The variables to specify when running the macro are the variable name
        (var), estimated mean for the first year (mu), coefficient of variation (cv), and
        standard deviation (std).  The model assumes the level of variation (cv and std) is the
        same both years.;

%macro glmpowery (var=, mu=, cv=, std=);

%do i=1 %to 5;

data tempyr;
   i=&i;
   if i=1 then diff=1.1;
   if i=2 then diff=1.2;
   if i=3 then diff=1.3;
   if i=4 then diff=1.4;
   if i=5 then diff=1.5;
   call symput("diff",trim(left(diff)));
   y1=&mu;
   y2=&mu * diff;

proc print data=tempyr;
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "normal - mean=&mu - cv=&cv - std=&std - diff=&diff";
run;

ods exclude all;

proc glmpower data=tempyr;
   title2 "normal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
   model y1 y2=;
   repeated year;
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std matrix('MyCorrs')=(0.2) corrs='MyCorrs';
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std matrix('MyCorrs')=(0.5) corrs='MyCorrs';
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std matrix('MyCorrs')=(0.8) corrs='MyCorrs';
   ods output output=outputyr;
run;

data outputyr&i;
   set outputyr;
   if dependent='year';
   if analysis='Power1' then rho=0.2;
   if analysis='Power2' then rho=0.5;
   if analysis='Power3' then rho=0.8;

ods exclude none;

proc print data=outputyr&i;
   title2 "normal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
```

```
run;

%end;

%mend glmpowery;
*************************************************************************************************;


*************************************************************************************************;
*** Proc glmpower macro - lognormal data
        This macro estimates the sample sizes needed to detect a 10, 20, 30, 40, and 50 percent
        difference between yearly means at a significance level of 0.05 and with 80 and 90
        percent power. In this macro, the model used for comparing years is a repeated measures
        model and assumes data are lognormally distributed. Sample sizes are estimated for a
        given mean and level of variation, and for three correlation levels (0.2, 0.5, and 0.8)
        between years. The variables to specify when running the macro are the variable name
        (var), estimated mean for the first year (mu), coefficient of variation (cv), and
        standard deviation (std). The model assumes the level of variation (cv and std) is the
        same both years. Mean and standard deviation are specified on a log scale.;

%macro glmpowerly (var=, mu=, cv=, std=);

%do i=1 %to 5;

data templyr;
   i=&i;
   if i=1 then diff=1.1;
   if i=2 then diff=1.2;
   if i=3 then diff=1.3;
   if i=4 then diff=1.4;
   if i=5 then diff=1.5;
   call symput("diff",trim(left(diff)));
   y1=&mu;
   y2=&mu + log(diff);

proc print data=templyr;
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "lognormal - mean=&mu - cv=&cv - std=&std - diff=&diff";
run;

ods exclude all;

proc glmpower data=templyr;
   title2 "lognormal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
   model y1 y2=;
   repeated year;
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std matrix('MyCorrs')=(0.2) corrs='MyCorrs';
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std matrix('MyCorrs')=(0.5) corrs='MyCorrs';
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std matrix('MyCorrs')=(0.8) corrs='MyCorrs';
   ods output output=outputlyr;
run;

data outputlyr&i;
   set outputlyr;
   if dependent='year';
   if analysis='Power1' then rho=0.2;
   if analysis='Power2' then rho=0.5;
   if analysis='Power3' then rho=0.8;

ods exclude none;

proc print data=outputlyr&i;
   title2 "lognormal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
run;

%end;

%mend glmpowerly;
*************************************************************************************************;
```

```
*************************************************************************************************;
*** Simulate beta distributed data.
        This macro simulates two years of correlated beta-distributed data for a savanna area.
        Data were simulated, at three correlation levels (0.2, 0.5, and 0.8), using a given mean
        and standard deviation and assuming a 0.05, 0.10, 0.15, 0.20, or 0.25 difference between
        the yearly means. For each combination of correlation and difference, datasets of
        various numbers of observations (sample sizes) were simulated. The sample sizes used
        varied with the difference between means to limit the number of datasets that needed to
        be created but yet have large enough sample sizes to determine where 80 and 90 percent
        power were achieved. A maximum sample size of 120 was used regardless of whether or not
        80 and 90 percent power were achieved. For each combination of correlation, difference,
        and sample size, a specified number of replicate datasets were simulated. The variables
        to specify when running the macro are the variable name (var), estimated mean for the
        first year (mu), standard deviation (std), and the number of replicate datasets to
        create (runs). The macro assumes the level of variation (std) is the same both years.;

%macro betay (var=, mu=, std=, runs=);

data yr1_beta;
  do run=1 to &runs;
    do rho=0.2 to 0.8 by 0.3;
      do d=1 to 5 by 1;
        if d=1 then do;  diff=0.05;  num=120;  b=4;  min=8;  end;
        if d=2 then do;  diff=0.10;  num=52;   b=2;  min=2;  end;
        if d=3 then do;  diff=0.15;  num=32;   b=1;  min=2;  end;
        if d=4 then do;  diff=0.20;  num=18;   b=1;  min=2;  end;
        if d=5 then do;  diff=0.25;  num=14;   b=1;  min=2;  end;
        do n=min to num by b;
          do plot=1 to n;
            a1 = (((&mu * (1 - &mu)) / (&std * &std)) - 1) * &mu;
            b1 = (((&mu * (1 - &mu)) / (&std * &std)) - 1) * (1 - &mu);
            a2 = ((((&mu - diff) * (1 - (&mu - diff))) / (&std * &std)) - 1) * (&mu - diff);
            b2 = ((((&mu - diff) * (1 - (&mu - diff))) / (&std * &std)) - 1) * (1 - (&mu - diff));
            x1 = rand("Normal",0,1);
            x2 = (rho*x1) + (sqrt((1-(rho**2)))*rand("Normal",0,1));
            z1 = cdf("Normal",x1);
            z2 = cdf("Normal",x2);
            y1 = quantile("Beta", z1, a1, b1);
            y2 = quantile("Beta", z2, a2, b2);
            output;
          end;
        end;
      end;
    end;
  end;
run;

proc sort data=yr1_beta;
   by run rho diff n plot;

proc transpose data=yr1_beta out=yearst(rename=col1=y) name=yr;
   by run rho diff n plot;
   var y1 y2;

data yr2_beta;
   set yearst;
   if yr='y1' then year=1;
   if yr='y2' then year=2;
   drop yr;

proc print data=yr2_beta (obs=50);
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "simulated beta data - mean=&mu - std=&std";

proc sort data=yr2_beta;
   by diff year;

run;

%mend betay;
```

```
*************************************************************************************************;


*************************************************************************************************;
*** Simulate Poisson distributed data.
        This macro simulates two years of correlated Poisson distributed data for a savanna area.
        Data were simulated, at three correlation levels (0.2, 0.5, and 0.8), using a given mean
        and assuming one of 5 difference levels between the yearly means. The differences used
        in the macro are computed as a function of the values 1, 2, 3, 4, and 5. For each
        combination of correlation and difference, datasets of various numbers of observations
        (sample sizes) were simulated. The sample sizes used varied with the difference between
        means to limit the number of datasets that needed to be created but yet have large
        enough sample sizes to determine where 80 and 90 percent power were achieved. A maximum
        sample size of 120 was used regardless of whether or not 80 and 90 percent power were
        achieved. For each combination of correlation, difference, and sample size, a specified
        number of replicate datasets were simulated. The variables to specify when running the
        macro are the variable name (var), estimated mean for the first year (mu), multiplier
        for computing the differences (dm) (e.g., if dm=0.5 then the differences used were 0.5,
        1, 1.5, 2, and 2.5), and the number of replicate datasets to create (runs).;

%macro county (var=, mu=, dm=, runs=);

data yr1_poisson;
   do run=1 to &runs;
      do rho=0.2 to 0.8 by 0.3;
         do d=1 to 5 by 1;
            if d=1 then do;  diff=d*&dm;  num=120;  b=4;  end;
            if d=2 then do;  diff=d*&dm;  num=91;   b=3;  end;
            if d=3 then do;  diff=d*&dm;  num=62;   b=2;  end;
            if d=4 then do;  diff=d*&dm;  num=34;   b=1;  end;
            if d=5 then do;  diff=d*&dm;  num=18;   b=1;  end;
            do n=4 to num by b;
               do plot=1 to n;
                  x1 = rand("Normal",0,1);
                  x2 = (rho*x1) + (sqrt((1-(rho**2)))*rand("Normal",0,1));
                  z1 = cdf("Normal",x1);
                  z2 = cdf("Normal",x2);
                  y1 = quantile("Poisson", z1, &mu);
                  y2 = quantile("Poisson", z2, &mu + diff);
                  output;
               end;
            end;
         end;
      end;
   end;
run;

proc sort data=yr1_poisson;
   by run rho diff n plot;

proc transpose data=yr1_poisson out=yearst(rename=col1=y) name=yr;
   by run rho diff n plot;
   var y1 y2;

data yr2_poisson;
   set yearst;
   if yr='y1' then year=1;
   if yr='y2' then year=2;
   drop yr;

proc print data=yr2_poisson (obs=50);
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "simulated Poisson data - mean=&mu";

proc sort data=yr2_poisson;
   by diff year;

run;

%mend county;
*************************************************************************************************;
```

```
*************************************************************************************************;
*** Simulate negative binomial distributed data.
        This macro simulates two years of correlated negative binomial distributed data for a
        savanna area. Data were simulated, at three correlation levels (0.2, 0.5, and 0.8),
        using a given mean and assuming one of 5 difference levels between the yearly means. The
        differences used in the macro are computed as a function of the values 1, 2, 3, 4, and
        5. For each combination of correlation and difference, datasets of various numbers of
        observations (sample sizes) were simulated. The sample sizes used varied with the
        difference between means to limit the number of datasets that needed to be created but
        yet have large enough sample sizes to determine where 80 and 90 percent power were
        achieved. A maximum sample size of 120 was used regardless of whether or not 80 and 90
        percent power were achieved. For each combination of correlation, difference, and sample
        size, a specified number of replicate datasets were simulated. The macro assumes the
        variance is a multiple of the mean each year. The variables to specify when running the
        macro are the variable name (var), estimated mean for the first year (mu), multiplier
        for computing the variance (m) (e.g., if m=2 then the variance is 2 times the mean),
        multiplier for computing the differences (dm) (e.g., if dm=1 then the differences used
        were 1, 2, 3, 4, and 5), minimum sample size (min), and the number of replicate data
        sets to create (runs).;

%macro nby (var=, mu=, m=, dm=, runs=, min=);

data yr1_negbin;
   do run = 1 to &runs;
      do rho = 0.2 to 0.8 by 0.3;
         do d = 1 to 5 by 1;
            diff = d*&dm;
            do n = &min to 120 by 4;
               do plot = 1 to n;
                  var1 = &m * &mu;
                  p1 = &mu / var1;
                  k1 = (&mu * &mu) / (var1 - &mu);
                  var2 = &m * (&mu + diff);
                  p2 = (&mu + diff) / var2;
                  k2 = ((&mu + diff) * (&mu + diff)) / (var2 - (&mu + diff));
                  x1 = rand("Normal",0,1);
                  x2 = (rho*x1) + (sqrt((1-(rho**2)))*rand("Normal",0,1));
                  z1 = cdf("Normal",x1);
                  z2 = cdf("Normal",x2);
                  _y1 = quantile("NegB", z1, p1, k1);
                  _y2 = quantile("NegB", z2, p2, k2);
                  y1=round(_y1,1);
                  y2=round(_y2,1);
                  output;
               end;
            end;
         end;
      end;
   end;
run;

proc sort data=yr1_negbin;
   by run rho diff n plot;

proc transpose data=yr1_negbin out=yearst(rename=col1=y) name=yr;
   by run rho diff n plot;
   var y1 y2;

data yr2_negbin;
   set yearst;
   if yr='y1' then year=1;
   if yr='y2' then year=2;
   drop yr;

proc print data=yr2_negbin (obs=50);
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "simulated negative binomial data - mean=&mu - m=&m";

proc sort data=yr2_negbin;
```

```
   by diff year;

run;

%mend nby;
*****************************************************************************************;


*****************************************************************************************;
*** Simulate binomial distributed data.
        This macro simulates two years of correlated binomial distributed data for a savanna
        area. Data were simulated, at three correlation levels (0.2, 0.5, and 0.8), using a
        given mean probability and assuming a 0.1, 0.2, 0.3, 0.4, or 0.5 difference between the
        yearly means. For each combination of correlation and difference, datasets of various
        numbers of observations (sample sizes) were simulated. The sample sizes used varied with
        the difference between means to limit the number of datasets that needed to be created
        but yet have large enough sample sizes to determine where 80 and 90 percent power were
        achieved. A maximum sample size of 120 was used regardless of whether or not 80 and 90
        percent power were achieved. For each combination of correlation, difference, and sample
        size, a specified number of replicate datasets were simulated. The variables to specify
        when running the macro are the variable name (var), estimated mean probability for the
        first year (p), and the number of replicate datasets to create (runs).;

%macro biny (var=, p=, runs=);

data yr1_binomial;
   do run=1 to &runs;
      do rho=0.2 to 0.8 by 0.3;
         do d=1 to 5 by 1;
            if d=1 then do;  diff=0.1;  num=120;  b=4;  end;
            if d=2 then do;  diff=0.2;  num=120;  b=4;  end;
            if d=3 then do;  diff=0.3;  num=63;   b=2;  end;
            if d=4 then do;  diff=0.4;  num=40;   b=1;  end;
            if d=5 then do;  diff=0.5;  num=34;   b=1;  end;
            do n=4 to num by b;
               do plot=1 to n;
                  p1 = &p;
                  p2 = &p + diff;
                  p1p2 = rho * sqrt(p1*(1-p1)*p2*(1-p2)) + (p1*p2);
                  t11=1-p1-p2+p1p2;
                  t12=p1-p1p2;
                  t21=p2-p1p2;
                  t22=1-t11-t12-t21;
                  cat=rand('TABLE', 1-p1-p2+p1p2, p1-p1p2, p2-p1p2);
                  if t11<0 or t12<0 or t21<0 or t22<0 then cat=.;
                  if cat=1 then do;  y1=0;  y2=0;  end;
                  if cat=2 then do;  y1=1;  y2=0;  end;
                  if cat=3 then do;  y1=0;  y2=1;  end;
                  if cat=4 then do;  y1=1;  y2=1;  end;
                  if cat=. then do;  y1=.;  y2=.;  end;
                  output;
               end;
            end;
         end;
      end;
   end;
run;

proc sort data=yr1_binomial;
   by run rho diff n plot;

proc transpose data=yr1_binomial out=yearst(rename=col1=y) name=yr;
   by run rho diff n plot;
   var y1 y2;

data yr2_binomial;
   set yearst;
   if yr='y1' then year=1;
   if yr='y2' then year=2;
   drop yr;
   if y=. then delete;
```

```
proc print data=yr2_binomial (obs=50);
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "simulated binomial data - p=&p";

proc sort data=yr2_binomial;
   by diff year;

run;

%mend biny;
*************************************************************************************************;


*************************************************************************************************;
*** Proc glimmix model.
        This macro runs the GLMM model to compare years within a savanna area. The model is a
        repeated measures model with year as a fixed effect and sample plot as a random effect.
        The glimmix procedure is used to run this model and most variables that need to be
        specified when running this macro are options within the glimmix procedure. After
        running all models, the macro computes the proportion of replicate runs that were
        significant at a significance level of 0.05 for each combination of correlation,
        difference, and sample size. This proportion is the power of the test at that sample
        size for each correlation and difference combination. The variables to specify when
        running the macro are the variable name (var), the distribution of the response variable
        (dist), the link function (link), important statistics for this run (stats) (this will
        depend on distribution and includes mean, standard deviation, and mean probability -
        these will just be used in the title for the output), model options that are specified
        as part of the proc glimmix statement (options), and other options which for these
        models will mainly be nonlinear options (other).;

%macro glimy (var=, dist=, link=, stats=, options=, other=);

ods listing exclude all;
ods results off;
options nonotes;

proc sort data=yr2_&dist;
   by run rho diff n plot;

proc glimmix data=yr2_&dist &options;
   by run rho diff n;
   &other;
   class year plot;
   model y=year / dist=&dist link=&link;
   random intercept / subject=plot;
   ods output tests3=glim1;
run;

data glim2;
   set glim1;
   if probf<=0.05 then sig=1;
   else sig=0;
run;

ods listing exclude none;
options notes;

proc summary data=glim2 nway;
   class rho diff n;
   var sig;
   output out=glim3(drop=_type_ _freq_) n=runs sum=;

data glim4;
   set glim3;
   power=sig/runs;

proc print data=glim4;
   title1 "&var - testing for a difference between 2 years for a savanna area";
   title2 "&dist - &stats - proc glimmix results";
run;
```

```
%mend glimy;
*************************************************************************************************;
```

```
*************************************************************************************************
*                                                                                               *
*     ***** Power analysis - Compare years.sas                                                  *
*                                                                                               *
*        *** Date first written:  4 May 2021                                                    *
*            Date last modified:  25 September 2021                                             *
*            Written by:  Deb Buhl                                                              *
*                                                                                               *
*  This program reads in the various macros needed for the power analyses for comparing years. *
*  Then for each variable, mean, and variance, the appropriate macros are run to estimate      *
*  sample sizes needed to detect specified differences between years with 80% and 90% power.   *
*                                                                                               *
*************************************************************************************************;


*************************************************************************************************;

*** Run code below to compile macros before running rest of this program.;

%include
   "c:\working files\savanna protocol\power analysis\power analysis - compare years macros.sas";
run;

   *** here is a list of the macros included in that program:
          %glmpowery (var=, mu=, cv=, std=)
               - glmpower to compare 2 years for a savanna area, normal data
          %glmpowerly (var=, mu=, cv=, std=)
               - glmpower to compare 2 years for a savanna area, lognormal data
          %betay (var=, mu=, std=, runs=)
               - simulate beta data for 2 years for a savanna area
          %county (var=, mu=, dm=, runs=)
               - simulate Poisson data for 2 years for a savanna area
          %nby (var=, mu=, m=, dm=, runs=, min=)
               - simulate negative binomial data for 2 years for a savanna area
          %biny (var=, p=, r=, d=, runs=)
               - simulate binomial data for 2 years for a savanna area
          %glimy (var=, dist=, link=, stats=, options=)
               - proc glimmix to compare 2 years for a savanna area

*************************************************************************************************;


*************************************************************************************************;
*** Tree density - Running power analysis to estimate sample sizes for comparing tree density
       between years for a savanna area. Analyses are run using glmpower procedure and assuming
       a normal distribution. Ran analyses with a mean of 255 and five variation levels.;

*** Using normal distribution;
proc datasets kill memtype=data; run;
%glmpowery (var=Trees/ha, mu=255, cv=30, std=76.5)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Trees/ha, mu=255, cv=45, std=114.75)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Trees/ha, mu=255, cv=60, std=153)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Trees/ha, mu=255, cv=75, std=191.25)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Trees/ha, mu=255, cv=90, std=229.5)
run;



*************************************************************************************************;
```

```
*** Basal area - Running power analysis to estimate sample sizes for comparing basal area
       between years for a savanna area. Analyses are run using glmpower procedure and assuming
       a normal distribution. Ran analyses with a mean of 18.4 and five variation levels.;

*** Using normal distribution;
proc datasets kill memtype=data; run;
%glmpowery (var=Basal Area, mu=18.4, cv=24, std=4.416)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Basal Area, mu=18.4, cv=33, std=6.072)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Basal Area, mu=18.4, cv=42, std=7.728)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Basal Area, mu=18.4, cv=51, std=9.384)
run;

proc datasets kill memtype=data; run;
%glmpowery (var=Basal Area, mu=18.4, cv=60, std=11.04)
run;



*************************************************************************************************;
*** DBH - Running power analysis to estimate sample sizes for comparing DBH between years for a
       savanna area. Analyses are run using glmpower procedure and assuming a lognormal
       distribution. Ran analyses with a mean of 3.91 and five variation levels.;

*** Using lognormal distribution - using mean and std of log transformed values;
proc datasets kill memtype=data; run;
%glmpowerly (var=DBH, mu=3.91, cv=3.1, std=0.12)
run;

proc datasets kill memtype=data; run;
%glmpowerly (var=DBH, mu=3.91, cv=4.9, std=0.19)
run;

proc datasets kill memtype=data; run;
%glmpowerly (var=DBH, mu=3.91, cv=6.6, std=0.26)
run;

proc datasets kill memtype=data; run;
%glmpowerly (var=DBH, mu=3.91, cv=8.4, std=0.33)
run;

proc datasets kill memtype=data; run;
%glmpowerly (var=DBH, mu=3.91, cv=10.2, std=0.40)
run;



*************************************************************************************************;
*** Canopy cover - Running power analysis to estimate sample sizes for comparing canopy cover
       between years for a savanna area. Analyses done by simulating data from a beta
       distribution and running GLMM models to compute power. Analyses run using three
       mean levels and three variation levels.;

*** Using beta distribution;
proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.72, std=0.09, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.72 - std=0.09,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.72, std=0.13, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.72 - std=0.13,
```

```
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.72, std=0.17, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.72 - std=0.17,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.78, std=0.09, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.78 - std=0.09,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.78, std=0.13, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.78 - std=0.13,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.78, std=0.17, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.78 - std=0.17,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.84, std=0.09, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.84 - std=0.09,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.84, std=0.13, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.84 - std=0.13,
        options=method=laplace)
run;

proc datasets kill memtype=data; run;
%betay (var=Canopy Cover, mu=0.84, std=0.17, runs=1000)
%glimy (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.84 - std=0.17,
        options=method=laplace)
run;



*************************************************************************************************;
*** Sapling counts - Running power analysis to estimate sample sizes for comparing sapling
       counts between years for a savanna area. Analyses done by simulating data from both
       Poisson and negative binomial distributions and running GLMM models to compute power.
       Analyses with a Poisson distribution are run using 3 mean values. Analyses with a
       negative binomial distribution are run using three mean levels and three variation
       levels.;

*** Using Poisson distribution;
proc datasets kill memtype=data; run;
%county (var=Sapling Counts, mu=2, dm=1, runs=1000)
%glimy (var=Sapling Counts, dist=Poisson, link=log, stats=mean=2,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%county (var=Sapling Counts, mu=6, dm=1, runs=1000)
%glimy (var=Sapling Counts, dist=Poisson, link=log, stats=mean=6,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%county (var=Sapling Counts, mu=10, dm=1, runs=1000)
%glimy (var=Sapling Counts, dist=Poisson, link=log, stats=mean=10,
```

```
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

*** Using negative binomial distribution;
proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=2, m=5, dm=1, runs=1000, min=8)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=2 - m=5,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=2, m=11, dm=1, runs=1000, min=12)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=2 - m=11,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=2, m=18, dm=1, runs=1000, min=12)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=2 - m=18,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=6, m=5, dm=1, runs=1000, min=12)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=6 - m=5,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=6, m=11, dm=1, runs=1000, min=8)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=6 - m=11,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=6, m=18, dm=1, runs=1000, min=8)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=6 - m=18,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=10, m=5, dm=1, runs=1000, min=12)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=10 - m=5,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=10, m=11, dm=1, runs=1000, min=8)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=10 - m=11,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Sapling Counts, mu=10, m=18, dm=1, runs=1000, min=8)
%glimy (var=Sapling Counts, dist=negbin, link=log, stats=mean=10 - m=18,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;



*************************************************************************************************;
*** Oak seedling counts - Running power analysis to estimate sample sizes for comparing oak
       seedling counts between years for a savanna area. Analyses done by simulating data from
       both Poisson and negative binomial distributions and running GLMM models to compute
       power. Analyses with a Poisson distribution are run using 3 mean values. Analyses with
       a negative binomial distribution are run using three mean levels and three variation
       levels.;

*** Using Poisson distribution;
proc datasets kill memtype=data; run;
%county (var=Oak Seedling Counts, mu=0.3, dm=0.5, runs=1000)
```

```
%glimy (var=Oak Seedling Counts, dist=Poisson, link=log, stats=mean=0.3,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%county (var=Oak Seedling Counts, mu=0.9, dm=0.5, runs=1000)
%glimy (var=Oak Seedling Counts, dist=Poisson, link=log, stats=mean=0.9,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%county (var=Oak Seedling Counts, mu=1.5, dm=0.5, runs=1000)
%glimy (var=Oak Seedling Counts, dist=Poisson, link=log, stats=mean=1.5,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

*** Using negative binomial distribution;
proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=0.3, m=2, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.3 - m=2,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=0.3, m=4, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.3 - m=4,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=0.3, m=6, dm=0.5, runs=1000, min=16)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.3 - m=6,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=0.9, m=2, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.9 - m=2,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=0.9, m=4, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.9 - m=4,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=0.9, m=6, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.9 - m=6,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=1.5, m=2, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=1.5 - m=2,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=1.5, m=4, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=1.5 - m=4,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;

proc datasets kill memtype=data; run;
%nby (var=Oak Seedling Counts, mu=1.5, m=6, dm=0.5, runs=1000, min=12)
%glimy (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=1.5 - m=6,
        options=method=laplace maxopt=300, other=nloptions tech=nrridg maxiter=300)
run;
```

```
*************************************************************************************************;
***Frequency from nested plots - Running power analysis to estimate sample sizes for comparing
       frequency from nested frequency plots between years for a savanna area. Analyses done by
       simulating data from a binomial distribution and running GLMM models to compute power.
       Analyses run using four mean probability levels.;

*** Using binomial distribution;
proc datasets kill memtype=data; run;
%biny (var=Frequency, p=0.1, runs=1000)
%glimy (var=Frequency, dist=binomial, link=logit, stats=p=0.1, options=method=laplace)
run;

proc datasets kill memtype=data; run;
%biny (var=Frequency, p=0.2, runs=1000)
%glimy (var=Frequency, dist=binomial, link=logit, stats=p=0.2, options=method=laplace)
run;

proc datasets kill memtype=data; run;
%biny (var=Frequency, p=0.3, runs=1000)
%glimy (var=Frequency, dist=binomial, link=logit, stats=p=0.3, options=method=laplace)
run;

proc datasets kill memtype=data; run;
%biny (var=Frequency, p=0.4, runs=1000)
%glimy (var=Frequency, dist=binomial, link=logit, stats=p=0.4, options=method=laplace)
run;



*************************************************************************************************;
```

```
*************************************************************************************************
*                                                                                               *
*     ***** Power analysis - Compare savannas macros.sas                                        *
*                                                                                               *
*        *** Date first written:  27 April 2021                                                 *
*            Date last modified:  25 September 2021                                             *
*            Written by:  Deb Buhl                                                              *
*                                                                                               *
*  This program contains macros needed for running power analyses on tests to compare 2         *
*  savanna areas.  There is a macro for running proc glmpower and several macros for            *
*  simulating data from various distributions (beta, Poisson, negative binomial, binomial,      *
*  and multinomial).  There is a macro for running a GLM model using proc glimmix to compare    *
*  savanna areas.                                                                               *
*                                                                                               *
*************************************************************************************************;


*************************************************************************************************;
*** Proc glmpower macro - normal data.
        This macro estimates the sample sizes needed to detect a 10, 20, 30, 40, and 50 percent
        difference between means from two savanna areas at a significance level of 0.05 and
        with 80 and 90 percent power. In this macro, the model used for comparing savanna areas
        is a single factor model and assumes data are normally distributed. Sample sizes are
        estimated for a given mean and level of variation. The variables to specify when
        running the macro are the variable name (var), estimated mean for the first savanna area
        (mu), coefficient of variation (cv), and standard deviation (std). The model assumes the
        level of variation (cv and std) is the same for both savanna areas.;

%macro glmpowers (var=, mu=, cv=, std=);

%do i=1 %to 5;

data tempsu;
   i=&i;
   if i=1 then diff=1.1;
   if i=2 then diff=1.2;
   if i=3 then diff=1.3;
   if i=4 then diff=1.4;
   if i=5 then diff=1.5;
   call symput("diff",trim(left(diff)));
   retain diff;
   savarea=1;   y=&mu;   output;
   savarea=2;   y=&mu * diff;   output;

proc print data=tempsu;
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "normal - mean=&mu - cv=&cv - std=&std - diff=&diff";
run;

ods exclude all;

proc glmpower data=tempsu;
   title2 "normal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
   class savarea;
   model y=savarea;
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std;
   ods output output=outputsu;
run;

data outputsu&i;
   set outputsu;
   nperarea=ntotal/2;

ods exclude none;

proc print data=outputsu&i;
   title2 "normal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
run;

%end;
```

```
%mend glmpowers;
*************************************************************************************************;


*************************************************************************************************;
*** Proc glmpower macro - lognormal data,
        This macro estimates the sample sizes needed to detect a 10, 20, 30, 40, and 50 percent
        difference between means from two savanna areas at a significance level of 0.05 and
        with 80 and 90 percent power. In this macro, the model used for comparing savanna areas
        is a single factor model and assumes data are lognormally distributed. Sample sizes are
        estimated for a given mean and level of variation. The variables to specify when running
        the macro are the variable name (var), estimated mean for the first savanna area (mu),
        coefficient of variation (cv), and standard deviation (std). The model assumes the
        level of variation (cv and std) is the same for both savanna areas.;

%macro glmpowerls (var=, mu=, cv=, std=);

%do i=1 %to 5;

data templsu;
   i=&i;
   if i=1 then diff=1.1;
   if i=2 then diff=1.2;
   if i=3 then diff=1.3;
   if i=4 then diff=1.4;
   if i=5 then diff=1.5;
   call symput("diff",trim(left(diff)));
   retain diff;
   savarea=1;   y=&mu;   output;
   savarea=2;   y=&mu + log(diff);   output;

proc print data=templsu;
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "lognormal - mean=&mu - cv=&cv - std=&std - diff=&diff";
run;

ods exclude all;

proc glmpower data=templsu;
   title2 "lognormal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
   class savarea;
   model y=savarea;
   power alpha=0.05 power=0.9 0.8 ntotal=. stddev=&std;
   ods output output=outputlsu;
run;

data outputlsu&i;
   set outputlsu;
   nperarea=ntotal/2;

ods exclude none;

proc print data=outputlsu&i;
   title2 "lognormal - mean=&mu - cv=&cv - std=&std - diff=&diff - proc glmpower results";
run;

%end;

%mend glmpowerls;
*************************************************************************************************;


*************************************************************************************************;
*** Simulate beta distributed data.
        This macro simulates beta-distributed data for two savanna areas. Data were simulated
        using a given mean and standard deviation and assuming a 0.05, 0.10, 0.15, 0.20, or 0.25
        difference between means from the two savanna areas. For each difference level, data
        sets of various numbers of observations (sample sizes) were simulated. The sample sizes
        used varied with the difference level to limit the number of datasets that needed to be
        created but yet have large enough sample sizes to determine where 80 and 90 percent
        power were achieved. A maximum sample size of 120 was used regardless of whether or not
```

```
        80 and 90 percent power were achieved. For each combination of difference level and
        sample size, a specified number of replicate datasets were simulated. The variables to
        specify when running the macro are the variable name (var), estimated mean for the first
        savanna area (mu), standard deviation (std), and the number of replicate datasets to
        create (runs). The macro assumes the level of variation (std) is the same for both
        savanna areas.;

%macro betas (var=, mu=, std=, runs=);

data su_beta;
  do run=1 to &runs;
    do d=1 to 5 by 1;
      if d=1 then do;  diff=0.05;  num=120;  b=4;  min=12;  end;
      if d=2 then do;  diff=0.10;  num=91;   b=3;  min=4;  end;
      if d=3 then do;  diff=0.15;  num=62;   b=2;  min=2;  end;
      if d=4 then do;  diff=0.20;  num=34;   b=1;  min=2;  end;
      if d=5 then do;  diff=0.25;  num=18;   b=1;  min=2;  end;
      do n=min to num by b;
        do savarea=1 to 2;
          do plot=1 to n;
            a1 = (((&mu * (1 - &mu)) / (&std * &std)) - 1) * &mu;
            b1 = (((&mu * (1 - &mu)) / (&std * &std)) - 1) * (1 - &mu);
            if savarea=1 then y = rand("Beta", a1, b1);
            a2 = ((((&mu - diff) * (1 - (&mu - diff))) / (&std * &std)) - 1) * (&mu - diff);
            b2 = ((((&mu - diff) * (1 - (&mu - diff))) / (&std * &std)) - 1) * (1 - (&mu - diff));
            if savarea=2 then y = rand("Beta", a2, b2);
            output;
          end;
        end;
      end;
    end;
  end;
run;

proc print data=su_beta (obs=50);
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "simulated beta data - mean=&mu - std=&std";

proc sort data=su_beta;
   by diff savarea;

run;

%mend betas;
*************************************************************************************************;


*************************************************************************************************;
*** Simulate Poisson distributed data.
        This macro simulates Poisson distributed data for two savanna areas. Data were simulated
        using a given mean and assuming one of 5 difference levels between means from the two
        savanna areas. The differences used in the macro are computed as a function of the values
        1, 2, 3, 4, and 5. For each difference level, datasets of various numbers of
        observations (sample sizes) were simulated. The sample sizes used varied with the
        difference level to limit the number of datasets that needed to be created but yet have
        large enough sample sizes to determine where 80 and 90 percent power were achieved. A
        maximum sample size of 120 was used regardless of whether or not 80 and 90 percent power
        were achieved. For each combination of difference level and sample size, a specified
        number of replicate datasets were simulated. The variables to specify when running the
        macro are the variable name (var), estimated mean for the first savanna area (mu),
        multiplier for computing the differences (dm) (e.g., if dm=0.5 then the differences used
        were 0.5, 1, 1.5, 2, and 2.5), and the number of replicate datasets to create (runs).;

%macro counts (var=, mu=, dm=, runs=);

data su_poisson;
   do run=1 to &runs;
      do d=1 to 5 by 1;
         if d=1 then do;  diff=d*&dm;  num=120;  b=4;  end;
         if d=2 then do;  diff=d*&dm;  num=91;   b=3;  end;
```

```
         if d=3 then do;  diff=d*&dm;  num=62;   b=2;  end;
         if d=4 then do;  diff=d*&dm;  num=34;   b=1;  end;
         if d=5 then do;  diff=d*&dm;  num=18;   b=1;  end;
         do n=4 to num by b;
            do savarea=1 to 2;
               do plot=1 to n;
                  if savarea=1 then y = rand("Poisson", &mu);
                  if savarea=2 then y = rand("Poisson", (&mu + diff));
                  output;
               end;
            end;
         end;
      end;
   end;
run;

proc print data=su_poisson (obs=50);
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "simulated Poisson data - mean=&mu";

proc sort data=su poisson;
   by diff savarea;

run;

%mend counts;
*************************************************************************************************;


*************************************************************************************************;
*** Simulate negative binomial distributed data.
        This macro simulates negative binomial distributed data for two savanna areas. Data were
        simulated using a given mean and assuming one of 5 difference levels between means from
        the two savanna areas. The differences used in the macro are computed as a function of
        the values 1, 2, 3, 4, and 5. For each difference level, datasets of various numbers of
        observations (sample sizes) were simulated. The sample sizes used varied with the
        difference level to limit the number of datasets that needed to be created but yet have
        large enough sample sizes to determine where 80 and 90 percent power were achieved. A
        maximum sample size of 120 was used regardless of whether or not 80 and 90 percent power
        were achieved. For each combination of difference level and sample size, a specified
        number of replicate datasets were simulated. The macro assumes the variance is a
        multiple of the mean for each savanna area. The variables to specify when running the
        macro are the variable name (var), estimated mean for the first savanna area (mu),
        multiplier for computing the variance (m) (e.g., if m=2 then the variance is 2 times the
        mean), multiplier for computing the differences (dm) (e.g., if dm=1 then the differences
        used were 1, 2, 3, 4, and 5), and the number of replicate datasets to create (runs).;

%macro nbs (var=, mu=, m=, dm=, runs=);

data su_negbin;
   do run=1 to &runs;
      do d=1 to 5 by 1;
         _diff = d * &dm;
         diff=round(_diff,0.1);
         do n=8 to 120 by 4;
            do savarea=1 to 2;
               do plot=1 to n;
                  var1 = &m * &mu;
                  p1 = &mu / var1;
                  k1 = (&mu * &mu) / (var1 - &mu);
                  if savarea=1 then _y = rand("NegBinomial", p1, k1);
                  var2 = &m * (&mu + diff);
                  p2 = (&mu + diff) / var2;
                  k2 = ((&mu + diff) * (&mu + diff)) / (var2 - (&mu + diff));
                  if savarea=2 then _y = rand("NegBinomial", p2, k2);
                  y=round(_y,1);
                 output;
               end;
            end;
         end;
      end;
```

```
   end;
run;

proc print data=su_negbin (obs=50);
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "simulated negative binomial data - mean=&mu - m=&m";

proc sort data=su_negbin;
   by diff savarea;

run;

%mend nbs;
*************************************************************************************************;


*************************************************************************************************;
*** Simulate binomial distributed data.
        This macro simulates binomial distributed data for two savanna areas. Data were
        simulated using a given mean probability and assuming a 0.1, 0.2, 0.3, 0.4, or 0.5
        difference between means from the two savanna areas. For each difference level, data
        sets of various numbers of observations (sample sizes) were simulated. The sample sizes
        used varied with the difference level to limit the number of datasets that needed to
        be created but yet have large enough sample sizes to determine where 80 and 90 percent
        power were achieved. A maximum sample size of 120 was used regardless of whether or not
        80 and 90 percent power were achieved. For each combination of difference level and
        sample size, a specified number of replicate datasets were simulated. The variables to
        specify when running the macro are the variable name (var), estimated mean probability
        for the first savanna area (p), and the number of replicate datasets to create (runs).;

%macro bins (var=, p=, runs=);

data su_binomial;
   do run=1 to &runs;
      do d=1 to 5 by 1;
         if d=1 then do;  diff=0.1;  num=120;  b=4;  end;
         if d=2 then do;  diff=0.2;  num=120;  b=4;  end;
         if d=3 then do;  diff=0.3;  num=63;   b=2;  end;
         if d=4 then do;  diff=0.4;  num=34;   b=1;  end;
         if d=5 then do;  diff=0.5;  num=24;   b=1;  end;
         do n=4 to num by b;
            do savarea=1 to 2;
               do plot=1 to n;
                  if savarea=1 then y = rand("Bernoulli", &p);
                  if savarea=2 then y = rand("Bernoulli", (&p + diff));
                  output;
               end;
            end;
         end;
      end;
   end;
run;

proc print data=su_binomial (obs=50);
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "simulated binomial data - p=&p";

proc sort data=su_binomial;
   by diff savarea;

run;

%mend bins;
*************************************************************************************************;


*************************************************************************************************;
*** Simulate ordered multinomial distributed data.
        This macro simulates ordered multinomial distributed data for two savanna areas. Data
        were simulated using given mean probabilities and assuming a 0.05, 0.1, or 0.15
        difference between the two savanna areas. Differences were applied by subtracting
```

```
        difference from first two categories and adding difference to last two categories for
        the second savanna area. For each difference level, datasets of various numbers of
        observations (sample sizes) were simulated. The sample sizes used varied with the
        difference level to limit the number of datasets that needed to be created but yet have
        large enough sample sizes to determine where 80 and 90 percent power were achieved. A
        maximum sample size of 120 was used regardless of whether or not 80 and 90 percent power
        were achieved. For each combination of difference level and sample size, a specified
        number of replicate datasets were simulated. The variables to specify when running the
        macro are the variable name (var), estimated mean probabilities for the first savanna
        area (p1, p2, p3, p4, and p5), and the number of replicate datasets to create (runs).;

%macro mults (var=, p1=, p2=, p3=, p4=, p5=, runs=);

data su_multinomial;
  do run=1 to &runs;
    do d=1 to 3 by 1;
      if d=1 then do;  diff=0.05;  num=120;  b=4;  end;
      if d=2 then do;  diff=0.10;  num=120;  b=4;  end;
      if d=3 then do;  diff=0.15;  num=120;  b=4;  end;
      do n=12 to num by b;
        do savarea=1 to 2;
          do plot=1 to n;
            if savarea=1 then y = rand("Table", &p1, &p2, &p3, &p4, &p5);
            if savarea=2 then
              y = rand("Table", (&p1 - diff), (&p2 - diff), &p3, (&p4 + diff), (&p5 + diff));
            output;
          end;
        end;
      end;
    end;
  end;
run;

proc print data=su_multinomial (obs=50);
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "simulated multinomial data - p1=&p1, p2=&p2, p3=&p3, p4=&p4, p5=&p5";

proc sort data=su_multinomial;
   by diff savarea;

run;

%mend mults;
*************************************************************************************************;


*************************************************************************************************;
*** Proc glimmix model.
        This macro runs the GLM model to compare means between savanna areas. The model is a
        single factor model with savanna area as the fixed effect. The glimmix procedure is used
        to run this model and most variables that need to be specified when running this macro
        are options within the glimmix procedure. After running all models, the macro computes
        the proportion of replicate runs that were significant at a significance level of 0.05
        for each combination of difference level and sample size. This proportion is the power
        of the test at that sample size and difference level. The variables to specify when
        running the macro are the variable name (var), the distribution of the response variable
        (dist), the link function (link), important statistics for this run (stats) (this will
        depend on distribution and includes mean, standard deviation, and mean probability -
        these will just be used in the title for the output), and model options that are
        specified as part of the proc glimmix statement (options).;

%macro glims (var=, dist=, link=, stats=, options=);

ods listing exclude all;
ods results off;
options nonotes;

proc sort data=su_&dist;
   by run diff n;

proc glimmix data=su_&dist &options;
```

```
   by run diff n;
   class savarea;
   model y=savarea / dist=&dist link=&link;
   ods output tests3=glims1;
run;

data glims2;
   set glims1;
   if probf<=0.05 then sig=1;
   else sig=0;
run;

ods listing exclude none;
options notes;

proc summary data=glims2 nway;
   class diff n;
   var sig;
   output out=glims3(drop=_type_ _freq_) n=runs sum=;

data glims4;
   set glims3;
   power=sig/runs;

proc print data=glims4;
   title1 "&var - testing for a difference between 2 savanna areas";
   title2 "&dist - &stats - proc glimmix results";
run;

%mend glims;
*************************************************************************************************;
```

```
*************************************************************************************************
*                                                                                               *
*     ***** Power analysis - Compare savannas.sas                                               *
*                                                                                               *
*        *** Date first written:  4 May 2021                                                    *
*            Date last modified:  25 September 2021                                             *
*            Written by:  Deb Buhl                                                              *
*                                                                                               *
*  This program reads in the various macros needed for the power analyses for comparing         *
*  savanna areas.  Then for each variable, mean, and variance, the appropriate macros are run  *
*  to estimate sample sizes needed to detect specified differences between savanna areas with  *
*  80% and 90% power.                                                                           *
*                                                                                               *
*************************************************************************************************;


*************************************************************************************************;

*** Run code below to compile macros before running rest of this program.;

%include
 "c:\working files\savanna protocol\power analysis\power analysis - compare savannas macros.sas";
run;

   *** here is a list of the macros included in that program:
          %glmpowers (var=, mu=, cv=, std=)
               - glmpower to compare 2 savannas for normal data
          %glmpowerls (var=, mu=, cv=, std=)
               - glmpower to compare 2 savannas for lognormal data
          %betas (var=, mu=, std=, runs=)
               - simulate beta data for 2 savanna areas
          %counts (var=, mu=, dm=, runs=)
               - simulate Poisson data for 2 savanna areas
          %nbs (var=, mu=, m=, dm=, runs=)
               - simulate negative binomial data for 2 savanna areas
          %bins (var=, p=, runs=)
               - simulate binomial data for 2 savanna areas
          %mults (var=, p1=, p2=, p3=, p4=, p5=, runs=)
               - simulate multinomial data for 2 savanna areas
          %glims (var=, dist=, link=, stats=, options=)
               - proc glimmix to compare 2 savanna areas

*************************************************************************************************;


*************************************************************************************************;
*** Trees/ha - Running power analysis to estimate sample sizes for comparing tree density
       between two savanna areas.  Analyses are run using glmpower procedure and assuming
       a normal distribution. Ran analyses with a mean of 255 and five variation levels.;

*** Using normal distribution;
proc datasets kill memtype=data; run;
%glmpowers (var=Trees/ha, mu=255, cv=30, std=76.5)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Trees/ha, mu=255, cv=45, std=114.75)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Trees/ha, mu=255, cv=60, std=153)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Trees/ha, mu=255, cv=75, std=191.25)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Trees/ha, mu=255, cv=90, std=229.5)
run;
```

```
*************************************************************************************************;
*** Basal area - Running power analysis to estimate sample sizes for comparing basal area
       between two savanna areas. Analyses are run using glmpower procedure and assuming
       a normal distribution. Ran analyses with a mean of 18.4 and five variation levels.;

*** Using normal distribution;
proc datasets kill memtype=data; run;
%glmpowers (var=Basal Area, mu=18.4, cv=24, std=4.416)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Basal Area, mu=18.4, cv=33, std=6.072)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Basal Area, mu=18.4, cv=42, std=7.728)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Basal Area, mu=18.4, cv=51, std=9.384)
run;

proc datasets kill memtype=data; run;
%glmpowers (var=Basal Area, mu=18.4, cv=60, std=11.04)
run;



*************************************************************************************************;
*** DBH - Running power analysis to estimate sample sizes for comparing DBH between two
       savanna areas. Analyses are run using glmpower procedure and assuming a lognormal
       distribution. Ran analyses with a mean of 3.91 and five variation levels.;

*** Using lognormal distribution - using mean and std of log transformed values;
proc datasets kill memtype=data; run;
%glmpowerls (var=DBH, mu=3.91, cv=3.1, std=0.12)
run;

proc datasets kill memtype=data; run;
%glmpowerls (var=DBH, mu=3.91, cv=4.9, std=0.19)
run;

proc datasets kill memtype=data; run;
%glmpowerls (var=DBH, mu=3.91, cv=6.6, std=0.26)
run;

proc datasets kill memtype=data; run;
%glmpowerls (var=DBH, mu=3.91, cv=8.4, std=0.33)
run;

proc datasets kill memtype=data; run;
%glmpowerls (var=DBH, mu=3.91, cv=10.2, std=0.40)
run;



*************************************************************************************************;
*** Canopy cover - Running power analysis to estimate sample sizes for comparing canopy cover
       between two savanna areas. Analyses done by simulating data from a beta distribution and
       running GLM models to compute power. Analyses run using three mean levels and three
       variation levels.;

*** Using beta distribution;
proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.72, std=0.09, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.72 - std=0.09)
run;

proc datasets kill memtype=data; run;
```

```
%betas (var=Canopy Cover, mu=0.72, std=0.13, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.72 - std=0.13)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.72, std=0.17, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.72 - std=0.17)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.78, std=0.09, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.78 - std=0.09)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.78, std=0.13, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.78 - std=0.13)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.78, std=0.17, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.78 - std=0.17)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.84, std=0.09, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.84 - std=0.09)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.84, std=0.13, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.84 - std=0.13)
run;

proc datasets kill memtype=data; run;
%betas (var=Canopy Cover, mu=0.84, std=0.17, runs=1000)
%glims (var=Canopy Cover, dist=beta, link=logit, stats=mean=0.84 - std=0.17)
run;



*************************************************************************************************;
*** Sapling counts - Running power analysis to estimate sample sizes for comparing sapling
       counts between two savanna areas. Analyses done by simulating data from both Poisson
       and negative binomial distributions and running GLM models to compute power. Analyses
       with a Poisson distribution are run using 3 mean values. Analyses with a negative
       binomial distribution are run using three mean levels and three variation levels.;

*** Using Poisson distribution;
proc datasets kill memtype=data; run;
%counts (var=Sapling Counts, mu=2, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=Poisson, link=log, stats=mean=2)
run;

proc datasets kill memtype=data; run;
%counts (var=Sapling Counts, mu=6, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=Poisson, link=log, stats=mean=6)
run;

proc datasets kill memtype=data; run;
%counts (var=Sapling Counts, mu=10, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=Poisson, link=log, stats=mean=10)
run;

*** Using negative binomial distribution;
proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=2, m=5, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=2 - m=5)
run;

proc datasets kill memtype=data; run;
```

```
%nbs (var=Sapling Counts, mu=2, m=11, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=2 - m=11)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=2, m=18, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=2 - m=18)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=6, m=5, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=6 - m=5)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=6, m=11, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=6 - m=11)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=6, m=18, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=6 - m=18)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=10, m=5, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=10 - m=5)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=10, m=11, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=10 - m=11)
run;

proc datasets kill memtype=data; run;
%nbs (var=Sapling Counts, mu=10, m=18, dm=1, runs=1000)
%glims (var=Sapling Counts, dist=negbin, link=log, stats=mean=10 - m=18)
run;



*************************************************************************************************;
*** Oak seedling counts - Running power analysis to estimate sample sizes for comparing oak
       seedling counts between two savanna areas. Analyses done by simulating data from both
       Poisson and negative binomial distributions and running GLM models to compute power.
       Analyses with a Poisson distribution are run using 3 mean values. Analyses with a
       negative binomial distribution are run using three mean levels and three variation
       levels.;

*** Using Poisson distribution;
proc datasets kill memtype=data; run;
%counts (var=Oak Seedling Counts, mu=0.3, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=Poisson, link=log, stats=mean=0.3)
run;

proc datasets kill memtype=data; run;
%counts (var=Oak Seedling Counts, mu=0.9, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=Poisson, link=log, stats=mean=0.9)
run;

proc datasets kill memtype=data; run;
%counts (var=Oak Seedling Counts, mu=1.5, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=Poisson, link=log, stats=mean=1.5)
run;

*** Using negative binomial distribution;
proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=0.3, m=2, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.3 - m=2)
run;
```

```
proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=0.3, m=4, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.3 - m=4)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=0.3, m=6, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.3 - m=6)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=0.9, m=2, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.9 - m=2)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=0.9, m=4, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.9 - m=4)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=0.9, m=6, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=0.9 - m=6)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=1.5, m=2, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=1.5 - m=2)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=1.5, m=4, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=1.5 - m=4)
run;

proc datasets kill memtype=data; run;
%nbs (var=Oak Seedling Counts, mu=1.5, m=6, dm=0.5, runs=1000)
%glims (var=Oak Seedling Counts, dist=negbin, link=log, stats=mean=1.5 - m=6)
run;



*************************************************************************************************;
***Frequency from nested plots - Running power analysis to estimate sample sizes for comparing
       frequency from nested frequency plots between two savanna areas. Analyses done by
       simulating data from a binomial distribution and running GLM models to compute power.
       Analyses were run using four mean probability levels.;

*** Using binomial distribution;
proc datasets kill memtype=data; run;
%bins (var=Frequency, p=0.1, runs=1000)
%glims (var=Frequency, dist=binomial, link=logit, stats=p=0.1)
run;

proc datasets kill memtype=data; run;
%bins (var=Frequency, p=0.2, runs=1000)
%glims (var=Frequency, dist=binomial, link=logit, stats=p=0.2)
run;

proc datasets kill memtype=data; run;
%bins (var=Frequency, p=0.3, runs=1000)
%glims (var=Frequency, dist=binomial, link=logit, stats=p=0.3)
run;

proc datasets kill memtype=data; run;
%bins (var=Frequency, p=0.4, runs=1000)
%glims (var=Frequency, dist=binomial, link=logit, stats=p=0.4)
run;
```

```
*************************************************************************************************;
*** Shrub-vine percent cover class - Running power analysis to estimate sample sizes for
       comparing shrub-vine percent cover class probabilities between two savanna areas.
       Analyses done by simulating data from a ordered multinomial distribution and running
       GLM models to compute power. Analyses were run using three mean probability levels.;

*** Using multinomial distribution;
proc datasets kill memtype=data; run;
%mults (var=Shrub-Vine Percent Cover Class, p1=0.78, p2=0.16, p3=0.04, p4=0.01, p5=0.01,
        runs=1000)
%glims (var=Shrub-Vine Percent Cover Class, dist=multinomial, link=cumlogit,
        stats=p1=0.78 p2=0.16 p3=0.04 p4=0.01 p5=0.01)
run;

proc datasets kill memtype=data; run;
%mults (var=Shrub-Vine Percent Cover Class, p1=0.15, p2=0.35, p3=0.25, p4=0.13, p5=0.12,
        runs=1000)
%glims (var=Shrub-Vine Percent Cover Class, dist=multinomial, link=cumlogit,
        stats=p1=0.15 p2=0.35 p3=0.25 p4=0.13 p5=0.12)
run;

proc datasets kill memtype=data; run;
%mults (var=Shrub-Vine Percent Cover Class, p1=0.2, p2=0.2, p3=0.2, p4=0.2, p5=0.2,
        runs=1000)
%glims (var=Shrub-Vine Percent Cover Class, dist=multinomial, link=cumlogit,
        stats=p1=0.2 p2=0.2 p3=0.2 p4=0.2 p5=0.2)
run;



*************************************************************************************************;
```

For more information about this publication, contact:
Director, USGS Northern Prairie Wildlife Research Center
8711 37th Street Southeast
Jamestown, ND 58401
701–253–5500

For additional information, visit: https://www.usgs.gov/centers/npwrc

Publishing support provided by the
Rolla Publishing Service Center

ISSN 2331-1258 (online)
https://doi.org/10.3133/ofr20221053