Lisa Siefker
Project 1D: Forest Ecology Data
GTDA 5402
SP25

INTRODUCTION & DESCRIPTION OF DATA

The Forest Ecology Data Set documents the frequency of 28 tree species in an old-growth forest. The forest was divided into 67 equal plots, and each plot was divided into equal subplots. Surveyors counted and recorded trees and their species occurring in four randomly selected sub-plots within each of the 67 plots. Each row of data represents one species, and each column represents one forest plot. For this analysis, I will consider only species in rows six (*C. ovata*) through 28 (*S. albidum*).

PROBLEM OF INTEREST

In this project, I will fit both the binomial and Poisson distributions to the species frequency data. For each distribution, I will create plots of the observed species counts and overlay the species counts expected under each distribution. I will visually examine the plots to determine which distribution best fits the data and then provide a table of parameter estimates for each species for the distribution that best fits the data. Identifying a distribution that fits the data will allow for statistical inferences about how each species might be distributed across the larger forest population. The well-fitted distribution can also be used to predict the likelihood of observing a particular species in a plot where data was not collected, and it allows for standardized comparisons of parameters across species.

In particular, I am interested in estimating the average number of occurrences of a species per unit area (i.e., plot). The average number of occurrences will help to identify common and rare species, which could inform planting strategies or identify invasive species. Analysis of average number of occurrences pre unit area could lead to additional research questions related to why a particular species is more or less common.

A second parameter of interest is variance of tree counts across plots. Variance may indicate a tendency for a species to either cluster or spread out. It also may be interesting to compare variance across species to identify differences between species. This information could inform planting strategies or lead to additional research questions.

RESULTS & DISCUSSION

Exploratory data analysis revealed a sample mean of 7.79 trees per plot and a sample variance of 61.65 across all species. However, an analysis of summary statistics for each species showed a range of sample means from 0 to 2.12 and a range of variances across species from 0 to 14.01. Four species in the data set did not have any trees in the surveyed subplots. (See **Appendix 1**). The number of trees per plot varied from a minimum of 0 to a maximum of 39. (See **Fig. 1** and **Fig. 2**)

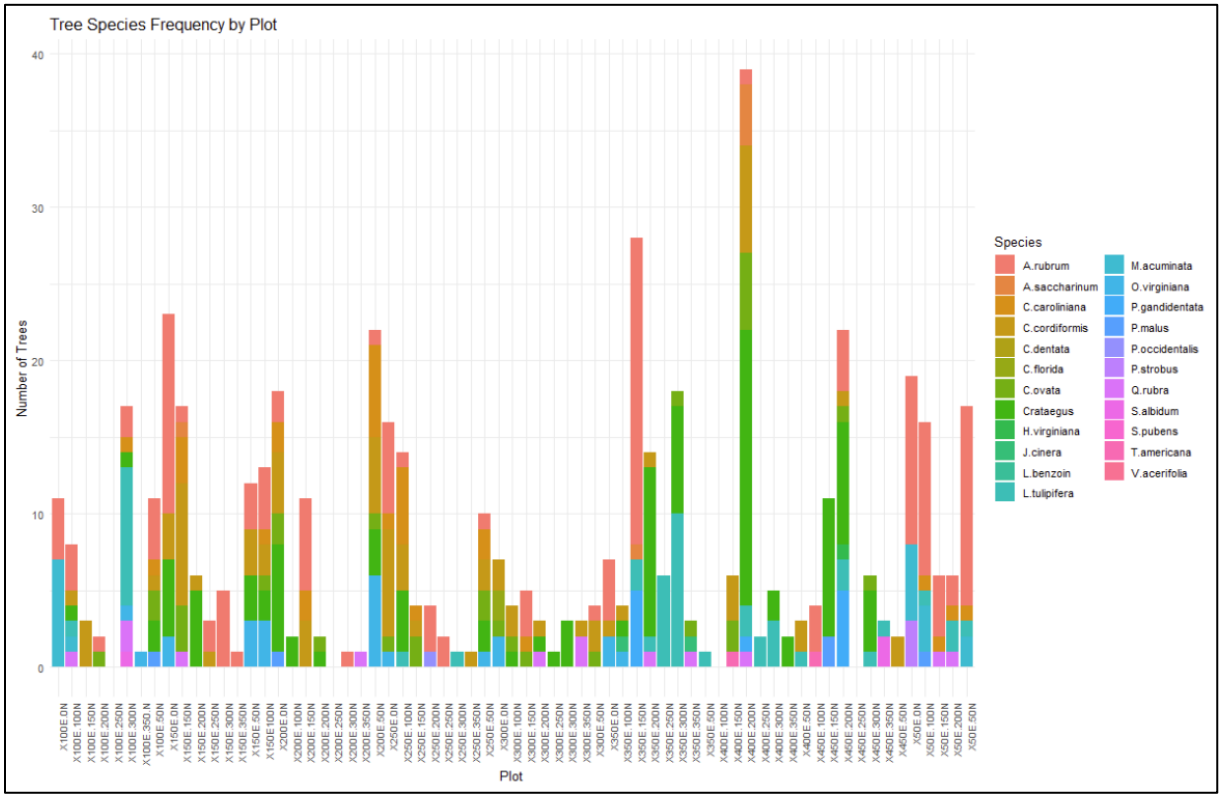


Figure 1. Stacked bar plot showing the count of trees by species in each plot.

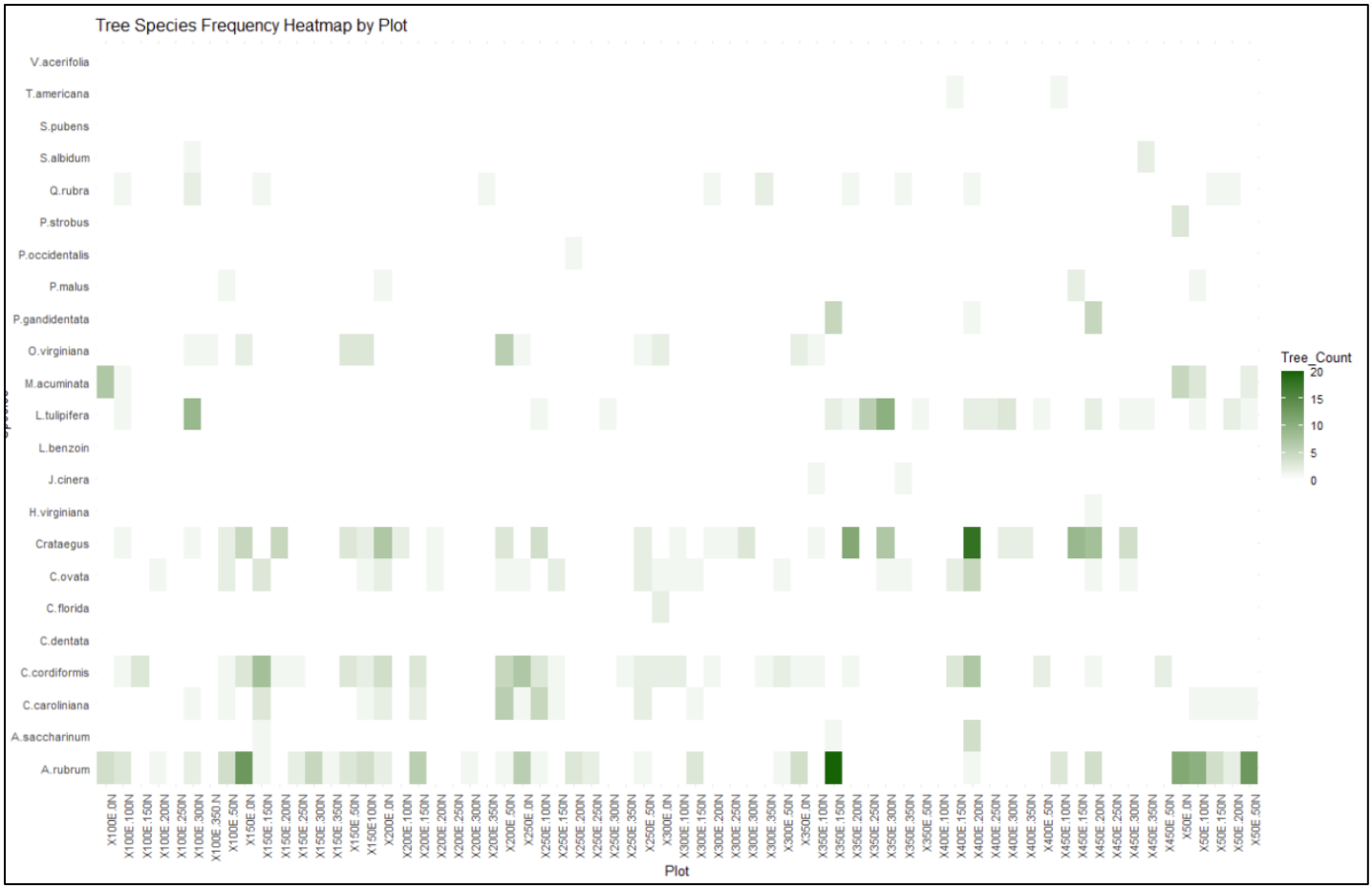


Figure 2. Heatmap showing the count of trees by species in each plot.

Fitting the Binomial Distribution

After exploring the data, I fit the binomial distribution to the frequency data for each species. Each plot was treated as an independent Bernoulli distribution, where failure indicates no trees of that species, and success indicates one or more trees of that species in the plot. The relevant parameters for the binomial distribution are n (the number of trials) and p (the probability of success on a single trial). For the forest data, the parameters are estimated as follows:

- n = 67 plots
- p = the probability of success for the species (# of successes/n)

To fit the binomial distribution to the data for each species, I plotted the observed frequencies from the data and overlayed the expected frequencies under the binomial distribution. A review of the plots shows large deviations between the observed and expected counts, which is evidence that the species data does not follow the binomial distribution. **(**See **Appendix #2)**

Fitting the Poisson Distribution

Next, I fit the Poisson distribution to the frequency data for each species. The Poisson distribution is used to model the number of "occurrences" in time or space. It has a single parameter, μ, which can be interpreted as the mean rate per unit area. The mean, μ, is equal to the variance for the Poisson distribution. This means that the spread or variability in the number of trees observed for each species is directly related to the average number of trees per plot. For the forest data, μ will be estimated with the sample mean of a species. To fit the Poisson distribution to the data for each species, I plotted the observed frequencies from the data and overlayed the expected values under the Poisson distribution. Poisson probabilities for each species are available in **Appendix 3**. A review of the plots shows that the Poisson distribution is a much better fit to the observed data than the binomial distribution.

A plot showing the Poisson distribution fit to each species is available in **Appendix 4**, and a more detailed view of two species is provided in **Fig. 3** and **Fig. 4.** If μ is small in a Poisson distribution, the distribution is more skewed, and if μ is large, the distribution approaches a normal distribution. The μ value was small for all species, and the faceted species plots show that the Poisson distribution is quite skewed for all species.

After fitting the Poisson distribution, I calculated the 95% confidence interval for the Poisson mean to understand the range of plausible values for the true population mean of each species. The widest confidence interval is for *A.rubrum* at 1.190, and the narrowest interval (for species with a mean > 0) is for

H.virginiana at .010. (See **Appendix 5**). A narrow confidence interval suggests a more precise estimate of the population mean, and a wider confidence interval indicates more uncertainty and larger variance in the data.

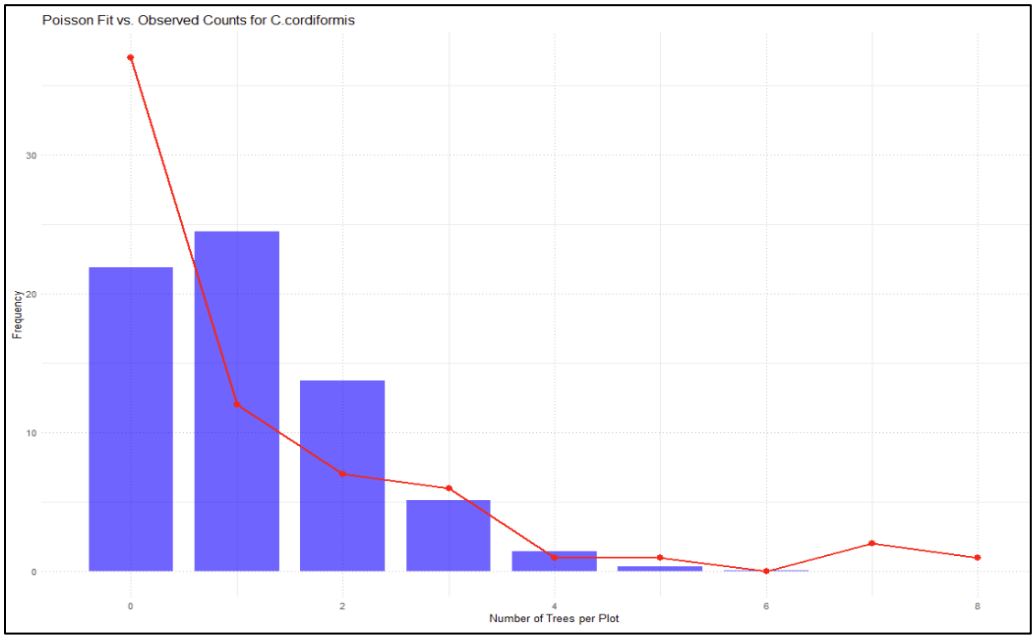


Figure 3. Plot showing observed values v. Poisson expected values for C.Cordiformis

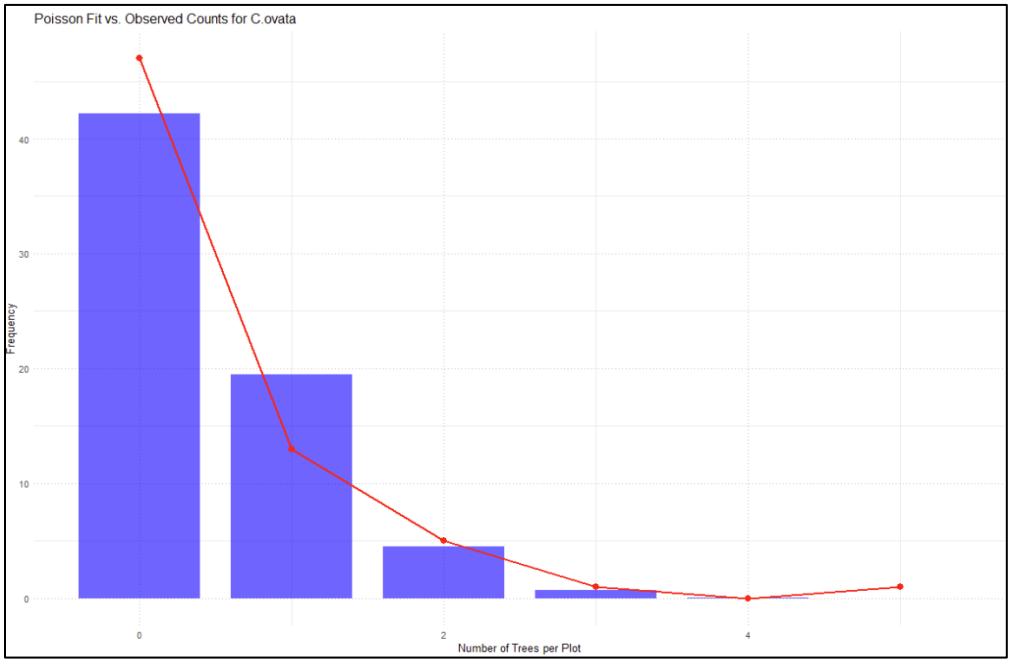


Figure 4. Plot showing observed values v. Poisson expected values for C.Ovata

CONCLUSIONS

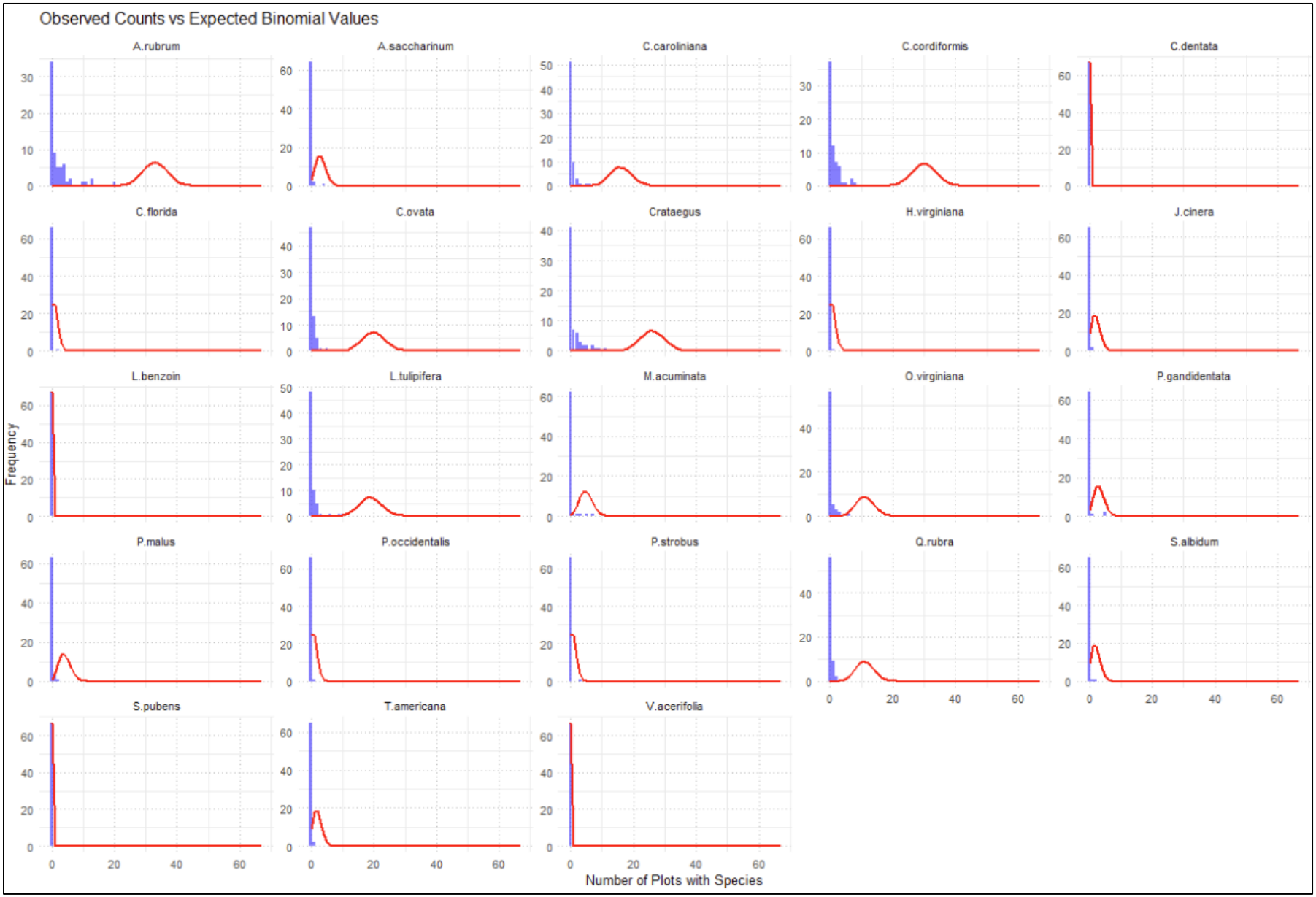
The Poisson distribution is reasonably well-aligned with observed values of species frequency. A good fit indicates that the Poisson distribution can provide insight into how each species might be distributed across the larger forest population, the likelihood of observing a particular species in a plot where data was not collected. This helps in estimating the probability of rare occurrences of a species in new plots. The Poisson distribution also allows for standardized comparisons of parameters across species.

APPENDIX 1: Summary Statistics by Species

	Species_Total <dbl>	Sample_Mean <dbl>	Std_Dev <dbl>	Variance <dbl>	Min <dbl>	Max <dbl>
A.rubrum	142	2.11940299	3.7437722	14.01582994	0	20
Crataegus	106	1.58208955	3.1629927	10.00452284	0	18
C.cordiformis	75	1.11940299	1.7966500	3.22795115	0	8
L.tulipifera	48	0.71641791	1.8324183	3.35775667	0	10
C.ovata	31	0.46268657	0.8932128	0.79782904	0	5
C.caroliniana	30	0.44776119	1.0910598	1.19041158	0	6
O.virginiana	23	0.34328358	0.9778182	0.95612845	0	6
M.acuminata	18	0.26865672	1.1225254	1.26006332	0	7
Q.rubra	13	0.19402985	0.4683564	0.21935776	0	2
P.gandidentata	11	0.16417910	0.8633447	0.74536409	0	5
A.saccharinum	6	0.08955224	0.5143794	0.26458616	0	4
P.malus	5	0.07462687	0.3168707	0.10040706	0	2
P.strobus	3	0.04477612	0.3665083	0.13432836	0	3
S.albidum	3	0.04477612	0.2715185	0.07372230	0	2
C.florida	2	0.02985075	0.2443389	0.05970149	0	2
J.cinera	2	0.02985075	0.1714598	0.02939846	0	1
T.americana	2	0.02985075	0.1714598	0.02939846	0	1
P.occidentalis	1	0.01492537	0.1221694	0.01492537	0	1
H.virginiana	1	0.01492537	0.1221694	0.01492537	0	1
L.benzoin	0	0.00000000	0.0000000	0.00000000	0	0
C.dentata	0	0.00000000	0.0000000	0.00000000	0	0
V.acerifolia	0	0.00000000	0.0000000	0.00000000	0	0
S.pubens	0	0.00000000	0.0000000	0.00000000	0	0

23 rows

APPENDIX 2: Observed Counts (blue) v. Expected Binomial Values (red line)

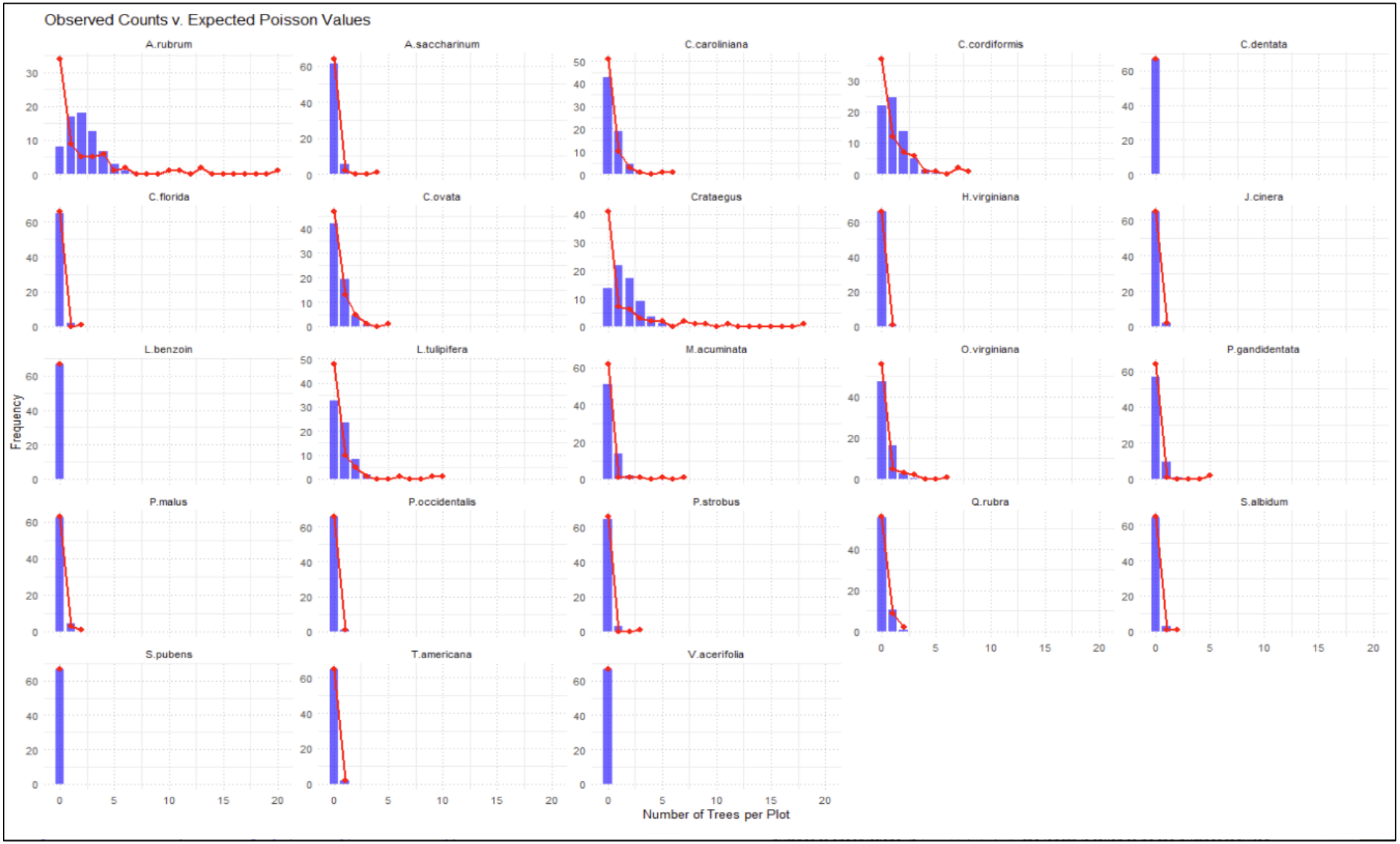


APPENDIX 3: Poisson probabilities for counts of each species

Species <chr>	Sample_Mean <dbl>	Poisson_Probs <chr>
C.ovata	0.46268657	0.6296, 0.2913, 0.0674, 0.0104, 0.0012, 1e-04
C.cordiformis	1.11940299	0.3265, 0.3655, 0.2045, 0.0763, 0.0214, 0.0048, 9e-04, 1e-04, 0
C.caroliniana	0.44776119	0.6391, 0.2861, 0.0641, 0.0096, 0.0011, 1e-04, 0
Q.rubra	0.19402985	0.8236, 0.1598, 0.0155
L.tulipifera	0.71641791	0.4885, 0.35, 0.1254, 0.0299, 0.0054, 8e-04, 1e-04, 0, 0, 0, 0
C.florida	0.02985075	0.9706, 0.029, 4e-04
Crataegus	1.58208955	0.2055, 0.3252, 0.2572, 0.1357, 0.0537, 0.017, 0.0045, 0.001, 2e-04, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
L.benzoin	0.00000000	1
A.rubrum	2.11940299	0.1201, 0.2545, 0.2697, 0.1906, 0.101, 0.0428, 0.0151, 0.0046, 0.0012, 3e-04, 1e-04, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
A.saccharinum	0.08955224	0.9143, 0.0819, 0.0037, 1e-04, 0
O.virginiana	0.34328358	0.7094, 0.2435, 0.0418, 0.0048, 4e-04, 0, 0
M.acuminata	0.26865672	0.7644, 0.2054, 0.0276, 0.0025, 2e-04, 0, 0, 0
P.malus	0.07462687	0.9281, 0.0693, 0.0026
C.dentata	0.00000000	1
P.gandidentata	0.16417910	0.8486, 0.1393, 0.0114, 6e-04, 0, 0
P.occidentalis	0.01492537	0.9852, 0.0147
V.acerifolia	0.00000000	1
J.cinera	0.02985075	0.9706, 0.029
T.americana	0.02985075	0.9706, 0.029
H.virginiana	0.01492537	0.9852, 0.0147
P.strobus	0.04477612	0.9562, 0.0428, 0.001, 0
S.pubens	0.00000000	1
S.albidum	0.04477612	0.9562, 0.0428, 0.001

23 rows

APPENDIX 4: Observed Counts (blue) v. Expected Poisson Values (red line)



APPENDIX 5: Confidence Intervals for the Poisson Mean

Species <chr>	Sample_Mean <dbl>	Poisson_Probs <list>	CI_95 <chr>
C.ovata	0.46268657	<dbl [6]>	(0.1847, 0.7407)
C.cordiformis	1.11940299	<dbl [9]>	(0.687, 1.5518)
C.caroliniana	0.44776119	<dbl [7]>	(0.1743, 0.7212)
Q.rubra	0.19402985	<dbl [3]>	(0.014, 0.3741)
L.tulipifera	0.71641791	<dbl [11]>	(0.3705, 1.0623)
C.florida	0.02985075	<dbl [3]>	(-0.0408, 0.1005)
Crataegus	1.58208955	<dbl [19]>	(1.068, 2.0961)
L.benzoin	0.00000000	<dbl [1]>	(0, 0)
A.rubrum	2.11940299	<dbl [21]>	(1.5244, 2.7144)
A.saccharinum	0.08955224	<dbl [5]>	(-0.0327, 0.2119)
O.virginiana	0.34328358	<dbl [7]>	(0.1038, 0.5827)
M.acuminata	0.26865672	<dbl [8]>	(0.0568, 0.4805)
P.malus	0.07462687	<dbl [3]>	(-0.037, 0.1863)
C.dentata	0.00000000	<dbl [1]>	(0, 0)
P.gandidentata	0.16417910	<dbl [6]>	(-0.0014, 0.3298)
P.occidentalis	0.01492537	<dbl [2]>	(-0.035, 0.0649)
V.acerifolia	0.00000000	<dbl [1]>	(0, 0)
J.cinera	0.02985075	<dbl [2]>	(-0.0408, 0.1005)
T.americana	0.02985075	<dbl [2]>	(-0.0408, 0.1005)
H.virginiana	0.01492537	<dbl [2]>	(-0.035, 0.0649)
P.strobus	0.04477612	<dbl [4]>	(-0.0417, 0.1313)
S.pubens	0.00000000	<dbl [1]>	(0, 0)
S.albidum	0.04477612	<dbl [3]>	(-0.0417, 0.1313)

23 rows