

Distribution Fitting 16. How Many Colors are actually in the Field?

Lorentz JÄNTSCHI¹⁾

¹⁾AcademicDirect, Muncii 103-105, Cluj-Napoca 400641 Cluj, Romania; lori@academicdirect.org.

Abstract. A study to compare different methods of estimation was conducted. The goal was to provide an estimate for the number of petal colors existing in the field by using a random sample of *Lycoris longituba* flowers taken from the field. Three methods of estimation were used to estimate the actual number of the colors in the field. Using the variance analysis, the maximum number of colors was obtained at 5% risk being in error. The study revealed that the estimation from observed sample (Fisher's method) agrees very well with third order Jackknife method of estimation.

Keywords: apparent abundance, diversity analysis, estimating methods, natural variability, petal color

INTRODUCTION

There are many methods estimating the abundance, some of them involving laborious calculations, such as the method of rarefaction. A review of the different approaches could be found in (Gotelli and Colwell, 2001). Unfortunately, most of the studies estimate species richness, but only some of them estimate qualitative traits.

The motivation of the study started from the desire to give an estimate for the number of colors of the flowers in the field based on a random sample. Thus, a series of estimating methods were adapted to be used on this concern and were applied on a data series taken from the literature.

The aim of the study was to find one or more alternatives to calculate the number of the colors of the flowers in the field based on a given random sample.

MATERIALS AND METHODS

Lycoris longituba is a cultivar of lily (*Amaryllidaceae*) family originated from China. Some authors consider the *Lycoris longituba* a distinct species (Hsu and Fan, 1974; Hsu *et al.*, 1994). Its flowers (Fig. 1) have many colors, being attractive to bees, butterflies and birds.

The data included in this study were taken from (He *et al.*, 2011). Tab. 1 contains the composition by color of a random sample of 44 flowers sampled randomly from a population in the natural distribution of this species in their natural habitat in China.

Tab. 1.

A random sample of *Lycoris longituba* from the field (according to He *et al.*, 2011)

Color	purple	red	yellow	orange	white
Flowers	14	10	10	9	1



Fig. 1. *Lycoris longituba* flower (adapted from Hsu and Fan, 1974)

First method used in this study was derived by Fisher in 1943 (Fisher, 1943) for estimating the abundance of the species. Under assumption of negative binomial distribution of the population subject to sampling, Fisher obtained a limiting form of the distribution excluding zero observations and applicable when a sample drawn from the population did not catch its entire diversity. Let C be the observed number of flower's colors and N the total number of flowers sampled. Then, according to (Fisher, 1943), C and N are in a relationship:

$$C - \alpha \cdot \ln\left(1 + \frac{N}{\alpha}\right) = 0 \quad (1)$$

with a unique solution ($\alpha = \alpha(C, N)$) with the meaning of richness in colors. Even more, under this assumption of an incomplete sample from negative binomial distribution, the variance of C for given N is:

$$\text{Var}(C) = \alpha \cdot \ln\left(\frac{2N + \alpha}{N + \alpha}\right) - \frac{\alpha^2 N}{(N + \alpha)^2} \quad (2)$$

Second applied method was the Bootstrap method. The Bootstrap method gives the estimates (Efron, 1979) from a repeated resampling with replacement:

- ÷ Give equal weight (of $1/n$ for a sample of size n) to each sampled observation;
- ÷ For each i from 1 to N (usually N from 50 to 200)
 - Extract (with replacement) a sample of (same) size n from the (initial) sample by using discrete uniform distribution and calculate the desired parameter (C) on it (let be c_i its value);
- ÷ The Bootstrap estimates for the parameter (C) and its variance ($\text{Var}(C)$) are:

$$C_B = \frac{1}{N} \sum_{i=1}^N c_i, \quad \text{Var}(C_B) = \frac{1}{N-1} \sum_{i=1}^N (C_B - c_i)^2 \quad (3)$$

Third applied method was the Jackknife method. The Jackknife method can be applied by removing systematically one or more observations from the sample (Quenouille, 1956). If only one observation is removed at a time, then n sub-samples of $n-1$ observations each (then the bias is of $O(1/n)$ order - see Quenouille, 1949) are obtained. If two observations are removed at a time, $n(n-1)/2$ sub-samples of $n-2$ observations each are obtained. If three

observations are removed at a time, $n(n-1)(n-2)/6$ sub-samples of $n-3$ observations each (and so on) are obtained. For groups up to size k , the bias is removed to order $O(1/n^k)$. Following algorithm gives the Jackknife estimates:

÷ For any k (here from 1 to 3) do For each i from 1 to $Combin(n,k)$

- Remove k observations from sample and calculate the desired parameter (C) (let be c_i its value);
- Compute $CJ_i = (n \cdot C - (n-k) \cdot c_i)/k$;

÷ The Jackknife estimates for the parameter and its variance are:

$$C_{jk} = \frac{1}{Combin(n, k)} \sum_{i=1}^N CJ_i, \quad Var(C_{jk}) = \frac{1}{n \cdot (Combin(n, k) - k)} \sum_{i=1}^N (C_{jk} - CJ_i)^2$$

(4)

RESULTS AND DISCUSSIONS

Assuming that the field is large enough (let's say over 10000 flowers) then normal distribution approximates well the sampling distribution of the number of colors for a given sample of flowers and any of the above mentioned methods can be applied to give the estimates of the desired parameter - number of the colors.

By applying Fisher's method, we need only the observed number of colors ($C = 5$, see Tab. 1) and the observed number of flowers ($N = 44$, see Tab. 1). Numerical solving of the Equation 1 gives $\alpha = 1.451889\dots$; introducing of the values in the Equation 2 gives a variance of 0.938 for $N = 44$ and a variance of 1.007 for $N \rightarrow \infty$.

Since the other methods are heavy computing ones, we will provide here only the results. Thus, Tab. 2 contains the estimates obtained with all applied methods.

Tab. 2.

Estimates for the number of colors

Method	Colors	Var (Colors)	Remarks
Fisher	5	1.01	small estimate; largest variance
Bootstrap	4.695	0.213	smallest estimate; smallest variance
Jackknife first order	5.977	0.955	large estimate; large variance
Jackknife second order	5.9545	0.201	large estimate; smallest variance
Jackknife third order	5.932	0.270	

It is expected that the Fisher's method to provide the largest variance. It uses smallest possible amount of information: only the number of flowers from the sample and the number of their colors. In order to provide a clear answer to the question from the title: "How many colors are actually in the field?" the inverse from the cumulative density function of the normal distribution at a 5% significance level should be calculated. Tab. 3 gives these estimates.

If the estimates are compared it can be observed that the Bootstrap estimates are underestimates and first and second order Jackknife are overestimates (relatively to the supposition of negative binomial distribution - let's say the exact method).

There is no a general agreement about which method of numerical estimation should be used. The third order Jackknife proved the closest to the Fisher's method, but this is not a general conclusion. Thus, in (Severiano *et al.*, 2011) authors obtained better results with Jackknife than with Bootstrap, and on the opposite, in (Smith *et al.*, 1986) and (Hellmann *et al.*, 1999) the authors obtained better results with Bootstrap than with Jackknife.

Tab. 3.

How many different colors may have the flowers in the field?

Method	Equation at 5% risk being in error	Solution	Answer
Fisher	$P\left(\frac{C-5}{1} \leq 1.96\right) = 0.975$	$C \leq 6.98$	$C_{\max} \sim 7$
Bootstrap	$P\left(\frac{C-4.695}{0.462} \leq 1.96\right) = 0.975$	$C \leq 5.60$	$C_{\max} \sim 6$
Bootstrap*	$P\left(\frac{C-4.695}{0.462} \leq 1.64\right) = 0.95$	$C \leq 5.45$	$C_{\max} \sim 6$
Jackknife first order	$P\left(\frac{C-5.977}{0.977} \leq 1.96\right) = 0.975$	$C \leq 7.89$	$C_{\max} \sim 8$
Jackknife second order	$P\left(\frac{C-5.9545}{0.6601} \leq 1.96\right) = 0.975$	$C \leq 7.23$	$C_{\max} \sim 8$
Jackknife third order	$P\left(\frac{C-5.932}{0.5194} \leq 1.96\right) = 0.975$	$C \leq 6.95$	$C_{\max} \sim 7$
* Computed according to (Efron and Tibshirani, 1993)			

The Fisher's method seems working very well in the investigated case - its result is on average of the others, and all results are very close to each other - and may be considered the golden test. Even more, the obtained results sustain the idea that the true distribution of the colors in the field is negative binomial. Let us note that the negative binomial distribution is often occur in nature and it comes from a convolution of Poisson and Gamma distributions where the mixing distribution of the Poisson rate is the Gamma distribution (Jäntschi *et al.*, 2011).

CONCLUSIONS

The analysis for the number of colors of *Lycoris longituba* flowers shown that Fisher's method estimating the abundance agrees very well at its 95% confidence superior limit with the abundance estimated with third order Jackknife method. All investigated methods - Fisher's, Bootstrap, Jackknife of first, second and third orders provide estimates close to each other. This result sustains the Fisher's idea (adapted here from number of species to number of colors) that the observations follow the negative binomial distribution.

Acknowledgments. The study was supported by POSDRU/89/1.5/S/62371 through a postdoctoral fellowship for L. Jäntschi. The sponsors of the study had no role in the study design, data collection, data analysis, data interpretation or writing of this article.

Special thanks to the effective management of POSDRU/89/1.5/S/62371 for the visit study at 'Floriade' exhibition in Venlo, 2012 (L. Costoiu).

Manuscript of this paper was deposited on ArXiv e-print service on May 19, 2012 and it appeared online since Tue, 22 May 2012 00:00:00 GMT (arXiv identifier: [1205.4306](https://arxiv.org/abs/1205.4306)).

REFERENCES

1. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7(1):1-26.

2. Efron, B. and R.J. Tibshirani (1993). Chapter 14. Of: An introduction to the bootstrap. Boca Raton: Chapman & Hall/CRC.
3. Fisher, R.A. (1943). Part 3. Theoretical distribution for the apparent abundance of different species. *J. Anim. Ecol.* 12:54-58.
4. Gotelli, N.J. and R.K. Colwell (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4(4):379-391.
5. He, Q., Y. Shen, M. Wang, M. Huang, R. Yang, S. Zhu, L. Wang, Y. Xu and R. Wu (2011). Natural Variation in Petal Color in *Lycoris longituba* Revealed by Anthocyanin Components. *PLoS ONE* 6(8):e22098.
6. Hellmann, J.J. and G.W. Fowler (1999). Bias, precision and accuracy of four measures of species richness. *Ecol. App.* 9:824-834.
7. Hsu, Y. and G.J. Fan (1974). A New Species of *Lycoris*. *Acta Phytotaxo Sinica* 12(3):299-302.
8. Hsu, P.S., S. Kurita, Z.S. Yu and J.Z. Lin (1994). Synopsis of the genus *Lycoris* (Amaryllidaceae). *Sida, Contrib Botan.* 16:301-331.
9. Jäntschi, L., S.D. Bolboacă, M.C. Bălan and R.E. Sestraş (2011). Distribution fitting 13. Analysis of independent, multiplicative effect of factors. Application to the effect of essential oils extracts from plant species on bacterial species. Application to the factors of antibacterial activity of plant species, *BUASVM. Anim. Sci. Biotechnol.*, 68(1-2):323-331.
10. Quenouille, M.H. (1949). Approximate tests of correlation in time-series. *J. R. Stat. Soc. B.* 11(1):68-84.
11. Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika* 43(3-4):353-360.
12. Severiano, A., J.A. Carriço, D.A. Robinson, M. Ramirez and F.R. Pinto (2011). Evaluation of Jackknife and Bootstrap for Defining Confidence Intervals for Pairwise Agreement Measures. *PLoS ONE* 6(5):e19539.
13. Smith, E.P., R.B. Genter and J. Cairns (1986). Confidence intervals for the similarity between algal communities. *Hydrobiol.* 139:237-245.