Using the β–binomial distribution to characterize forest health[1]

S.J. Zarnoch, R.L. Anderson, and R.M. Sheffield

Abstract: The β–binomial distribution is suggested as a model for describing and analyzing the dichotomous data obtained from programs monitoring the health of forests in the United States. Maximum likelihood estimation of the parameters is given as well as asymptotic likelihood ratio tests. The procedure is illustrated with data on dogwood anthracnose infection (caused by *Discula destructiva*) in the southeastern United States. The parameter estimates have important biological interpretation, and tests of hypotheses are more meaningful than traditional statistical analyses. The value of a modeling approach to dichotomous data analysis is emphasized.

Résumé : La distribution β–binomiale est proposée comme modèle pour décrire et analyser les données dichotomiques qui proviennent des programmes de surveillance de la santé des arbres dans le pays. La valeur des paramètres qui a le maximum de probabilité est donnée de même que les tests de ratio de probabilité asymptotique. La procédure est illustrée avec les données d'une infection par l'anthracnose du cournouiller (causée par *Discula destructiva*) dans le sud-est des États-Unis. Les valeurs estimées des paramètres ont une importante signification biologique et les tests d'hypothèse ont plus de signification que les analyses statistiques traditionnelles. La valeur de cette approche qui utilise la modélisation est soulignée dans le cas de l'analyse de données dichotomiques.

[Traduit par la Rédaction]

Introduction

The recent interest in the effects of atmospheric deposition and global climate change on forests has led many government agencies and landowners to observe forest health during their periodic inventories of forest resources. In addition it has led to the creation of national programs to monitor forest health, such as the National Forest Health Monitoring Program. Many of the variables observed in these programs are dichotomous (binary). For example, trees are often classified as dead or alive due to a certain cause or simply as damaged or undamaged due to, say, ozone. A typical sampling design usually consists of cluster sampling where a random or systematic sample of plots is selected across the forest type and within each plot a cluster of several trees, shrubs, organisms, etc. is selected for observation.

The estimation of proportions can be performed using the typical cluster sampling formula (Cochran 1977), which is based on a ratio of two variables. However, a ratio of two variables has a complex distribution, the estimator is skewed and usually slightly biased, and little insight is obtained into the deviation of the response from pure binomial variation. If one simply ignores the clusters and calculates a proportion that is inappropriate but often done, the true standard error is usually underestimated (Kish 1957).

In testing hypotheses about proportions, the analyses are not necessarily straightforward. Usually a transformation is performed on the proportions and typical normal theory tests are used. However, if the proportions are based on different sample sizes, the variances are heterogeneous and a weighting procedure is required. In addition, if the normality assumption is questionable, a nonparametric analysis may have to be employed. Haseman and Kupper (1979) detail the many alternative approaches to analyzing proportions.

Recently, the β–binomial distribution has been used to study dichotomous variables under a cluster sampling design (Williams 1975; Kupper and Haseman 1978; Haseman and Kupper 1979). Within each cluster of size n, the number of individuals affected by a certain condition is assumed to have a binomial distribution with parameters n and p, where p is the probability that a given individual is affected. If p is constant across the population, then an observed frequency distribution of a sample of clusters should follow the typical binomial distribution. However, this is not the case when p varies from cluster to cluster according to the β distribution, which then yields the β–binomial model. In this situation, the tails of the observed frequency distribution are heavier than expected based on

Received April 7, 1994. Accepted October 1, 1994.

S.J. Zarnoch. USDA Forest Service, Southern Research Station, Athens, GA 30602, U.S.A.
R.L. Anderson. USDA Forest Service, Southern Region, Asheville, NC 28802, U.S.A.
R.M. Sheffield. USDA Forest Service, Southern Research Station, Asheville, NC 28804, U.S.A.

[1] The use of trade or firm names in this publication is for reader information and does not imply endorsement by the U.S. Department of Agriculture of any product or service.

the binomial distribution. Thus, under the β–binomial model the typical binomial variation as well as the extra-binomial variation are considered.

In comparing proportions, particularly to detect a change due to some environmental factor, it is not only important to compare the mean proportions but also the distribution of the proportions. For instance, two populations may have identical proportions of dead trees but one may exhibit a more uniform spatial pattern of dead trees over the population, while the other may have great variability. The former should be well represented by the binomial distribution (a special case of the β–binomial), while the latter conforms to the β–binomial. The reason for these differences in spatial patterns may be due to contagion of the disease itself or to specific environmental factors that are more conducive to the disease in one area and not in the other. This spatial pattern should be important information for the specialist who is dealing with forest health issues.

Alternative models to the β–binomial have also been proposed for proportions. Kupper and Haseman (1978) presented the correlated binomial model, which ignores the intercluster variation but specifically considers the pairwise correlations among the individuals in a cluster. Another model considered is the multiplicative generalized binomial (Altham 1978) based on a symmetric joint distribution. Although the β–binomial allows for positive association between individuals in a cluster, the correlated binomial and the multiplicative generalized binomial allow for positive and negative association. However, in comparison studies between these models, no clear-cut advantages were evident (Altham 1978; Haseman and Kupper 1979; Paul 1982). It has also been suggested that maximization of the likelihood function under the β–binomial is mathematically more tractable. Thus, for simplicity, we have considered only the β–binomial model but encourage investigation into the others.

Our objectives here are to describe the β–binomial distribution for use in forest health analysis and to outline the estimation of the parameters and tests of hypotheses. Throughout, data on dogwood anthracnose (caused by *Discula destructiva*), a disease of flowering dogwood (*Cornus florida* L.), are used to illustrate the utility of the β–binomial model.

Materials and methods

The β–binomial distribution

Let n be the number of trees observed on a plot, x be the number of trees observed on a plot that have the specified health condition under study, and p be the proportion of trees on a plot that have the specified health condition. Assuming that x has the binomial distribution defined as

$$P(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \qquad \begin{matrix} 0 \le x \le n \\ 0 < p < 1 \end{matrix}$$

then an estimator for p is

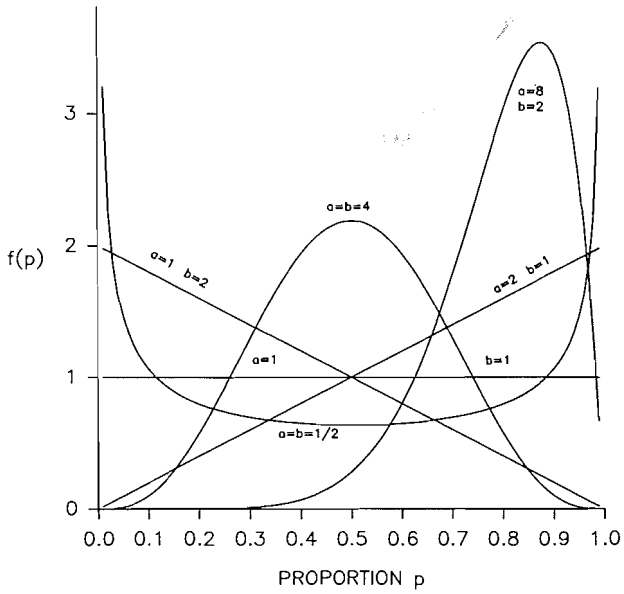
$$\hat{p} = \frac{x}{n}$$

Given a forest type of considerable size, say covering several hundred thousand acres, a forest health survey may consist of m plots that are sampled for the health condition. If the proportion p is the same throughout the entire forest type, then only binomial variation is present. However, it is possible that a plant disease may not exert its effect uniformly across the forest type because of the epidemiology of the disease and (or) the spatial variability over the environment. Thus, p may vary from plot to plot over the forest type, adding extra binomial variation. Assume that each p comes from the β distribution defined as

$$f(p) = \frac{\Gamma(a + b)}{\Gamma(a)\,\Gamma(b)} p^{a-1} (1 - p)^{b-1}, \qquad \begin{matrix} 0 < p < 1 \\ a > 0 \\ b > 0 \end{matrix}$$

where $\Gamma(z)$ is the gamma function evaluated at z. Figure 1 shows the β distribution $f(p)$ for various values of a and b. As both parameters approach zero, the proportion becomes concentrated at the tails of the distribution, which implies that most plots will either have no trees affected ($p = 0$) or have all trees affected ($p = 1$) with the health condition. As the parameters increase toward infinity, the distribution of the proportion becomes concentrated at a single constant that yields similar p across all plots, thus conforming to the pure binomial model. Because of the flexibility of the β distribution, a large range of intermediate conditions is possible.

Fig. 1. The β distribution for various values of the parameters a and b.



The joint distribution of x and p is simply the product of the binomial and β. Thus, the β–binomial is the marginal distribution of x, defined as

$$P(x) = \int_0^1 \binom{n}{x} p^x (1 - p)^{n-x} \frac{\Gamma(a + b)}{\Gamma(a)\,\Gamma(b)} \times p^{a-1}(1 - p)^{b-1} dp$$

$$= \binom{n}{x} \frac{B(a + x,\ n + b - x)}{B(a,\ b)} \qquad x = 0, 1, 2, \ldots, n$$

$$= 0, \qquad \text{elsewhere}$$

where $B(r, s)$ is the β function evaluated at the arguments r and s. The expectation of x is

$$E(x) = \frac{na}{a + b}$$

and the variance is

$$V(x) = \frac{nab(n + a + b)}{(a + b)^2 (1 + a + b)}$$

It is more meaningful to reparameterize (Griffiths 1973) to

$$\mu = \frac{a}{a + b}, \qquad 0 < \mu < 1$$

$$\theta = \frac{1}{a + b} \times \theta > 0$$

where μ is the expectation of the β distribution, that is, the mean proportion, and θ determines the shape of the β distribution. Note that when θ approaches zero (a and (or) b approach infinity) the β–binomial converts to the pure binomial, and when θ increases (a and (or) b approach zero) the proportions are concentrated in the tails of the distribution. This characteristic is quite useful because the magnitude of departure from pure binomial variation for a population can be quantified. The reparameterized β–binomial distribution is now defined as

$$P(x) = \binom{n}{x} \frac{B\left(x + \frac{\mu}{\theta}, n - x + \frac{1 - \mu}{\theta}\right)}{B\left(\frac{\mu}{\theta}, \frac{1 - \mu}{\theta}\right)} \qquad x = 0, 1, 2, \ldots, n$$

$$= 0, \qquad \text{elsewhere}$$

Estimation and hypotheses testing

The preferred method for estimating the parameters of the β–binomial distribution is maximum likelihood. The likelihood function is the product of the $P(x_i)$ based on each plot i, $i = 1, 2, \ldots, m$. Hence, the natural logarithm of the likelihood function is

$$\ln L(\mu,\theta) = \ln \prod_{i=1}^{m} P(x_i) = \ln \prod_{i=1}^{m} \left[\binom{n_i}{x_i} \frac{B\left(x_i + \frac{\mu}{\theta}, n_i - x_i + \frac{1 - \mu}{\theta}\right)}{B\left(\frac{\mu}{\theta}, \frac{1 - \mu}{\theta}\right)} \right]$$

$$= \sum_{i=1}^{m} \left[\ln \binom{n_i}{x_i} + \sum_{r=0}^{x_i - 1} \ln(\mu + r\theta) + \sum_{r=0}^{n_i - x_i - 1} \ln(1 - \mu + r\theta) - \sum_{r=0}^{n_i - 1} \ln(1 + r\theta) \right]$$

The log likelihood function could be maximized by differentiating with respect to μ and θ and solving a system of nonlinear equations. However, the simplex method of Nelder and Mead (1965) is quite useful for maximizing the log likelihood directly and is easily programmed in FORTRAN. Moment estimates, based on the sample mean and variance, can be used as initial starting solutions.

The log likelihood function can be used for various asymptotic tests of hypotheses (Williams 1975). For instance, say it is desired to test for homogeneity of two populations; that is, the proportion of affected trees is the same in both populations with identical variation in the proportions. This leads to the hypothesis

$$H_0: \mu_1 = \mu_2$$
$$\theta_1 = \theta_2$$

versus

$$H_1: \mu_1 \neq \mu_2$$
$$\theta_1 \neq \theta_2$$

Let L_0 be the maximum value of the log likelihood under H_0 and L_1 be the maximum value of the log likelihood under H_1. For this hypothesis test, L_0 can be obtained by maximizing the log likelihood when the observations from both populations are combined. This results in estimating only two parameters, a common μ and θ, which satisfies the constraints under H_0. Conversely, L_1 is obtained by summing the individual maximum log likelihoods obtained from separate maximizations for each population. This allows for no constraints on the parameter estimates as specified by H_1 and results in four parameters being estimated, μ_1, μ_2, θ_1, and θ_2. Note that $-\infty < L_0 \leq L_1 \leq 0$, since each log likelihood is the logarithm of a value between zero and one. The test statistic is $2(L_1 - L_0)$, which is compared with the upper percentage point of the χ^2 distribution with two degrees of freedom.

More specific hypotheses are also possible to test with the likelihood ratio approach, and they may be more powerful than the general test for homogeneity specified previously. For instance, say one wishes to test for a difference in the mean proportions for two populations given a common θ; that is

$$H_0: \mu_1 = \mu_2$$
$$\theta_1 = \theta_2$$

versus

$$H_1: \mu_1 \neq \mu_2$$
$$\theta_1 = \theta_2$$

L_0 is obtained as before (two parameters estimated), while L_1 is obtained by maximizing the log likelihood under the constraints of H_1, which results in the estimation of three parameters. The test statistic is $2(L_1 - L_0)$, which is compared with the upper percentage point of the χ^2 distribution with one degree of freedom. Other one degree of freedom tests are possible and are outlined in Table 1.

Results

Proportion of dogwoods infected with dogwood anthracnose

Dogwood anthracnose was first reported on flowering dogwood in New York and Connecticut in 1978. Since then

Table 1. Specific asymptotic likelihood ratio tests for various hypotheses about two populations, where the test statistic $2(L_1 - L_0)$ is compared with the upper percentage point of the χ^2 distribution with one degree of freedom.

Hypotheses	Maximum log likelihood[a]	Parameters estimated
	Different μ_j assuming common θ	
H_0: $\mu_1 = \mu_2$, $\theta_1 = \theta_2$	$L_0 = \max[\ln L_1(\mu, \theta) + \ln L_2(\mu, \theta)]$	μ, θ
H_1: $\mu_1 \neq \mu_2$, $\theta_1 = \theta_2$	$L_1 = \max[\ln L_1(\mu_1, \theta) + \ln L_2(\mu_2, \theta)]$	μ_1, μ_2, θ
	Different μ_j assuming different θ_j	
H_0: $\mu_1 = \mu_2$, $\theta_1 \neq \theta_2$	$L_0 = \max[\ln L_1(\mu, \theta_1) + \ln L_2(\mu, \theta_2)]$	μ, θ_1, θ_2
H_1: $\mu_1 \neq \mu_2$, $\theta_1 \neq \theta_2$	$L_1 = \max[\ln L_1(\mu_1, \theta_1) + \ln L_2(\mu_2, \theta_2)]$	μ_1, μ_2, θ_1, θ_2
	Different θ_j assuming common μ	
H_0: $\theta_1 = \theta_2$, $\mu_1 = \mu_2$	$L_0 = \max[\ln L_1(\mu, \theta) + \ln L_2(\mu, \theta)]$	μ, θ
H_1: $\theta_1 \neq \theta_2$, $\mu_1 = \mu_2$	$L_1 = \max[\ln L_1(\mu, \theta_1) + \ln L_2(\mu, \theta_2)]$	μ, θ_1, θ_2
	Different θ_j assuming different μ_j	
H_0: $\theta_1 = \theta_2$, $\mu_1 \neq \mu_2$	$L_0 = \max[\ln L_1(\mu_1, \theta) + \ln L_2(\mu_2, \theta)]$	μ_1, μ_2, θ
H_1: $\theta_1 \neq \theta_2$, $\mu_1 \neq \mu_2$	$L_1 = \max[\ln L_1(\mu_1, \theta_1) + \ln L_2(\mu_2, \theta_2)]$	μ_1, μ_2, θ_1, θ_2

[a]Ln $L_j(\mu_j, \theta_j)$ is the log likelihood function for population j, $j = 1, 2$.

it has spread south to Georgia (Anderson et al. 1990; USDA Forest Service 1988). Infection of trees is favored by cool, wet weather, and symptoms include leaf spots, leaf mortality, twig dieback, cankers, lower branch dieback, and tree mortality. Mortality often occurs 3–5 years after the initial infection. Very small trees die in 1 to 2 years. Under the direction of the USDA Forest Service, Forest Health, Southern Region, the Dogwood Anthracnose Impact Assessment Program was begun in the Southeast to help monitor the spread of the disease (USDA Forest Service 1991). A network of 210 permanent plots was established on a grid across the Southeast in 1988 and monitored annually to 1992. A plot usually consisted of 10 permanently marked dogwood trees whose states of health were individually assessed. The last 2 years of data (1990–1991) are used as an example of the application of the β–binomial model because they had the largest number of plots with dogwoods.

Table 2 shows the distribution of the observed data on a plot basis, along with the expected data, under binomial and β–binomial models. Chi-square tests indicate a poor fit to the binomial distribution. In both years the distribution is easily rejected at the $P = 0.05$ level. Hence, one can conclude that dogwoods are not being attacked by the disease uniformly across the Southeast. On the other hand, the β–binomial provides a very good fit to the frequency data and is not rejected at the $P = 0.05$ level. For 1990, the estimated parameters were $\hat{\mu} = 0.1542$ and $\hat{\theta} = 0.5324$. Figure 2 illustrates the associated β distribution $f(p)$, which indicates that most plots had a low proportion of trees infected and the frequency diminished monotonically as this proportion approached one. However, in 1991 the estimated parameters were $\hat{\mu} = 0.3000$ and $\hat{\theta} = 1.0087$. Figure 2 shows that this $f(p)$ is approaching a u-shaped distribution in which the proportion of trees infected on a plot is either low or high. The mean proportion has increased substantially. Also, $\hat{\theta}$ has increased, which indicates more variability across the region. There are now several plots with all 10 trees infected; before, there were none. The likelihood ratio test for homogeneity was used to test the hypothesis that there has been a change in the distributions from 1990 to 1991. The validity of this test is questionable because the data for each year are not independent, but it is used only to illustrate the test procedure. The results gave $L_0 = -1297.5132$, $L_1 = -598.8333 - 687.3730 = -1286.2063$, and thus, $2(L_1 - L_0) = 22.6138$. Comparing this to the χ^2 with two degrees of freedom gives a probability level of $P = 0.0000$, which rejects the hypothesis of homogeneity. Thus, there has been a change in the health status of dogwood trees in the Southeast with respect to dogwood anthracnose. To further examine this situation, specific one degree of freedom tests were performed (Table 3). The results indicate that there has been a shift in the mean proportion, μ_j, assuming either a common or different θ_j for both years. Tests for a change in variability, θ_j, were significant only when assuming

Table 2. Results from the Dogwood Anthracnose Impact Assessment Program for 1990 and 1991 and χ^2 goodness of fit tests for the binomial and β–binomial models.

	1990			1991		
Infected trees	Observed	Binomial expected	β–Binomial expected	Observed	Binomial expected	β–Binomial expected
0	92	28.5	89.2	63	3.9	61.8
1	15	55.3	24.4	20	17.6	18.9
2	16	48.3	14.8	8	35.7	12.7
3	13	25.0	10.5	9	42.8	10.1
4	6	8.5	8.0	4	33.8	8.7
5	9	2.0	6.2	12	18.3	7.9
6	9	0.3	4.9	6	6.9	7.4
7	5	0.0	3.8	13	1.8	7.2
8	2	0.0	2.9	7	0.3	7.4
9	1	0.0	2.1	12	0.0	8.0
10	0	0.0	1.2	7	0.0	10.7
Total	168	167.9	168.0	161	161.1	160.8
X^2, computed		443.2	12.0		1621.1	14.9
df		4	8		6	8
$\chi^2_{0.05}$, table		9.49	15.5		12.6	15.5

Note: For simplicity, only plots with exactly 10 dogwood trees were used in this analysis. Expected cell frequencies less than 1 were combined for the χ^2 tests.

different μ_j's. Thus, it appears that there has been a general change in the entire distribution of infected dogwoods between the years.

Proportion of dogwood mortality due to dogwood anthracnose

A survey was conducted by Forest Health (Southern Region) and Forest Inventory and Analysis (Southern Research Station) to determine the proportion of anthracnose-caused dogwood mortality in Forest Inventory and Analysis plots. Forest Inventory and Analysis identified 126 permanent sample plots in western North Carolina that had one or more dead dogwood trees. In the fall of 1991, a representative sample consisting of 39 plots was selected and each dead dogwood was examined to determine whether it had died from dogwood anthracnose or not. Presence of cankers and systematic epicormic shoots was considered indicative of anthracnose. Recently killed trees and those with some living tissue were checked in the laboratory to confirm the presence of *D. destructiva*. This procedure was repeated in Virginia in the spring and summer of 1992 on 20 additional plots. The results indicate that North Carolina had many plots with most or no dead dogwoods infected with *D. destructiva*, while Virginia had very few such plots (Table 4).

In this example, the n_i are variable and, hence, the standard χ^2 goodness of fit test for the binomial distribution is not applicable. Tarone (1979) developed three $C(\alpha)$ tests for goodness of fit to the binomial distribution that are asymptotically optimal against the β–binomial, the correlated binomial, and the multiplicative binomial. Although we have focused on the β–binomial model, all three tests were performed to illustrate the value of using a modeling approach to analyzing proportions instead of the typical binomial distribution. The binomial was rejected at the $P = 0.001$ level for all tests in favor of the other models. Since all test statistics were so large, no preference could be found for any one of the three alternatives; thus, all three would probably give similar inferences. Hence, our selection of the β–binomial is justifiable.

The β–binomial distribution was fitted separately by state and yielded parameter estimates of $\hat{\mu} = 0.3616$ and $\hat{\theta} = 1.0351$ for North Carolina and $\hat{\mu} = 0.4282$ and $\hat{\theta} = 0.1843$ for Virginia. Plots of these two distributions are shown in Fig. 3 and reflect quite opposite patterns of infection. North Carolina appears to exhibit a much higher degree of variability across the plots than Virginia. The likelihood ratio test was used to determine whether the distributions were homogeneous. The results gave $L_0 = -364.0647$, $L_1 = -177.9318 - 180.4073 = -358.3391$, and thus, $2(L_1 - L_0) = 11.4512$. Comparing this to the χ^2 with two degrees of freedom gives a probability level of $P = 0.0033$, which concludes that the two distributions differ. Specific one degree of freedom tests indicate no significant differences in the mean proportions μ_j, but a highly significant difference in variability θ_j assuming either common or differing μ_j's (Table 3). The biological interpretation for this difference is unknown, but by using the β–binomial model this distinction was easily observed. It is speculated that the percent of trees killed by dogwood anthracnose in Virginia may have been more uniform because the disease has been in the area longer, giving all sites an equal chance to be affected. In North Carolina, it is speculated that the disease has only recently become established and all sites have not had an equal chance to be affected. Many areas have not been affected, and others

Table 3. Specific tests for various hypotheses based on the test statistic $2(L_1 - L_0)$, which is compared with the upper percentage point of the χ^2 distribution with one degree of freedom.

	1990–1991 Impact Assessment example		North Carolina – Virginia example	
	$2(L_1 - L_0)$	P-value	$2(L_1 - L_0)$	P-value
Different μ_j assuming common θ				
H_0: $\mu_1 = \mu_2$, $\theta_1 = \theta_2$	15.0450	0.0001	1.0732	0.3002
H_1: $\mu_1 \neq \mu_2$, $\theta_1 = \theta_2$				
Different μ_j assuming different θ_j				
H_0: $\mu_1 = \mu_2$, $\theta_1 \neq \theta_2$	21.2486	0.0000	0.6924	0.4054
H_1: $\mu_1 \neq \mu_2$, $\theta_1 \neq \theta_2$				
Different θ_j assuming common μ				
H_0: $\theta_1 = \theta_2$, $\mu_1 = \mu_2$	1.3652	0.2426	10.7588	0.0010
H_1: $\theta_1 \neq \theta_2$, $\mu_1 = \mu_2$				
Different θ_j assuming different μ_j				
H_0: $\theta_1 = \theta_2$, $\mu_1 \neq \mu_2$	7.5688	0.0059	10.3780	0.0013
H_1: $\theta_1 \neq \theta_2$, $\mu_1 \neq \mu_2$				

that only recently became infected have had little opportunity to inflict mortality on the dogwood trees.

It is interesting to perform other standard statistical analyses on this data and compare the results to the β–binomial likelihood ratio test. Analysis of variance (ANOVA) on the raw data is one such test often used to detect location differences in proportions between two groups. Since each proportion is based on a variable sample size, one may also consider weighting each observation by the reciprocal of the variance of the proportion, for instance with $n/p(1 - p)$ or simply n. Another option often used is the arcsine square root proportion transformation. If there is question about the validity of the normality assumption, one may use nonparametric tests like the Wilcoxon rank–sum, the median, the Van der Waerden, and the Savage. Another promising method recently developed for comparing proportions based on clustered data has been given by Rao and Scott (1992). The results from all location difference tests were far from significant, except for the median test (Table 5). This is not surprising, since the mean proportions for North Carolina and Virginia were quite similar, and thus, these tests are at a distinct disadvantage when compared with the β–binomial likelihood ratio test, which could also detect shifts in variability. The objective here was not to critically compare these tests but to illustrate that commonly used location difference tests are inadequate for detecting changes in the distribution of a proportion. Two more general tests useful for detecting a change in the total distribution are the Kolmogorov–Smirnov and the Kuiper. These distribution tests fared better when compared with the β–binomial likelihood ratio test, but both had lower significance levels (Table 5).

Fig. 2. The β distribution for the probability of dogwood anthracnose (p) fitted to the 1990 and 1991 Dogwood Impact Assessment Program plot data.

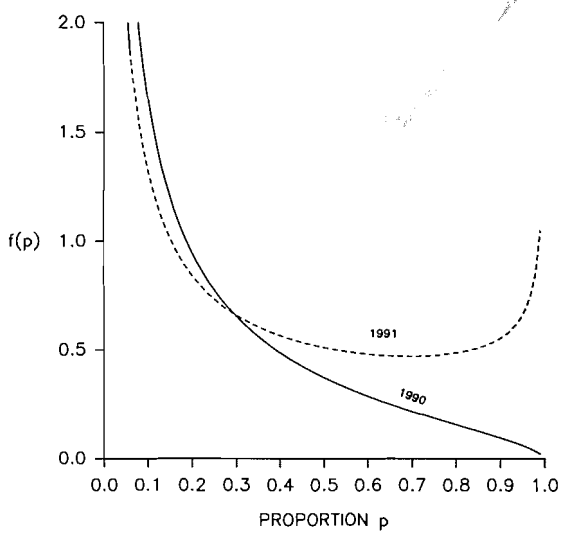
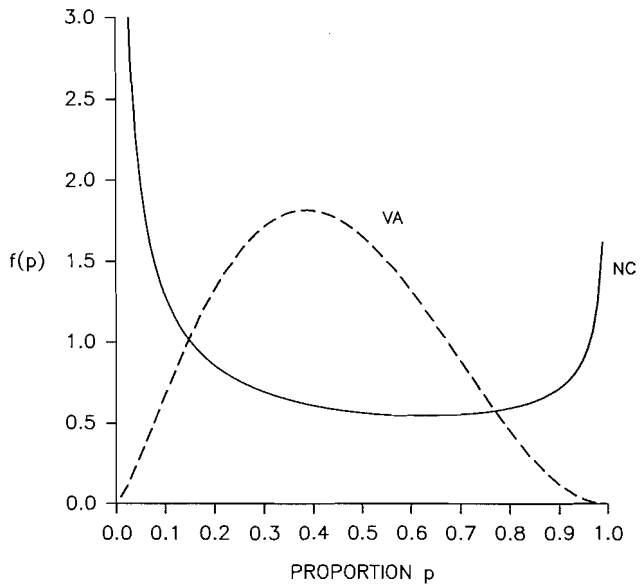


Fig. 3. The β distribution for the probability of a dead dogwood being infected by dogwood anthracnose (p) in North Carolina and Virginia.



Discussion and conclusions

The β–binomial distribution was quite useful in analyzing proportional data associated with forest health issues,

Table 4. Data on a plot basis from the North Carolina – Virginia Dogwood Anthracnose Survey.

North Carolina						Virginia		
n	*x*	*p*	*n*	*x*	*p*	*n*	*x*	*p*
3	2	0.67	7	0	0.00	9	4	0.44
2	0	0.00	9	3	0.33	20	18	0.90
5	0	0.00	13	12	0.92	12	0	0.00
7	5	0.71	20	6	0.30	7	2	0.29
14	8	0.57	10	2	0.20	20	11	0.55
3	3	1.00	14	4	0.29	10	4	0.40
1	0	0.00	8	0	0.00	18	13	0.72
8	8	1.00	4	0	0.00	14	8	0.57
1	1	1.00	3	0	0.00	14	2	0.14
10	10	1.00	6	6	1.00	20	14	0.70
7	7	1.00	15	10	0.67	15	3	0.20
15	0	0.00	7	2	0.29	5	1	0.20
11	1	0.09	20	13	0.65	12	5	0.42
7	0	0.00	8	0	0.00	20	7	0.35
10	4	0.40	6	0	0.00	9	2	0.22
20	3	0.15	15	7	0.47	16	7	0.44
9	0	0.00	5	0	0.00	20	12	0.60
9	0	0.00	14	8	0.57	15	7	0.47
9	5	0.56	7	1	0.14	10	1	0.10
4	0	0.00				12	7	0.58

Note: *n*, number of dead dogwood trees on a plot; *x*, number of dead dogwood trees infected by dogwood anthracnose; $p = x/n$, proportion of dead dogwood trees infected by dogwood anthracnose.

Table 5. Alternative statistical tests for the North Carolina – Virginia Dogwood Anthracnose Survey.

Analysis procedure	*P*-value
Location differences	
ANOVA	0.5458
ANOVA, $w = n/p(1 - p)$[a]	0.6213
ANOVA, $w = n$	0.2996
ANOVA, arcsine($p^{0.5}$)	0.4410
Wilcoxon rank–sum test	0.2471
Median test	0.0838
Van der Waerden test	0.3253
Savage test	0.7741
Rao–Scott test	0.3056
Distributional differences	
Kolmogorov–Smirnov test	0.0647
Kuiper test	0.0130

[a]Only 37 observations out of 59 could be used due to *p* being zero.

but its use is not limited to such situations. It should be applicable to most dichotomous data dealing with issues such as seed germination, tree survival, insect pest attacks, and disease screening.

The traditional analyses of proportions, such as ANOVA and nonparametrics that emphasize changes in location only, should be avoided. A more general and preferable approach is to detect changes in the distribution of proportions, which could be detected with the Kolmogorov–Smirnov and Kuiper tests. However, the β–binomial provides a modeling approach to analyzing proportions with parameters that are biologically interpretable. In addition, asymptotic likelihood ratio tests are available and appear competitive to the distributional tests. Although considerable computing effort is required to obtain the maximum likelihood estimates and likelihood ratio tests, it is insignificant in comparison to the effort required to collect such data. A FORTRAN program to perform these calculations is available from the authors.

By using the β–binomial with a modeling approach, we were able to detect different distributional patterns of dogwood anthracnose infection over time and space. Results led to speculation about the cause of the patterns and the possibility of further hypothesis formulation, research, and sampling efforts. If the traditional approaches such as ANOVA had been used, such trends would have gone undetected, perhaps until a much more distinct pattern had evolved at a later date. In forest health situations, the objective is to detect and understand such changes as soon as possible so that corrective measures can be employed.

The β–binomial distribution is quite flexible and should fit many biological databases. In particular, it includes the binomial as a special case ($\theta = 0$) and the negative binomial as a limiting form. Unlike traditional analyses, it handles unequal sample sizes without problem and requires no special weighting or transformations of the data.

Acknowledgments

The authors thank the Associate Editor and the reviewers for their comments and suggestions. Appreciation is also extended to Linda Watson, Southern Research Station, for her efforts in preparing this manuscript.

References

Altham, P.M.E. 1978. Two generalizations of the binomial distribution. Appl. Stat. **27**(2): 162–167.

Anderson, R.L., Knighten, J.L., and Dowsett, S.E. 1990. Dogwood anthracnose: a southeastern United States perspective. *In* 23rd Annual Tennessee Nursery Short Course, 14–16 Feb. 1990, Maxwell House Hotel, Nashville. Tennessee Nursery Industry, Nashville. pp. 242–253.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. John Wiley & Sons, New York.

Griffiths, D.A. 1973. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, **29**: 637–648.

Haseman, J.K., and Kupper, L.L. 1979. Analysis of dichotomous response data from certain toxicological experiments. Biometrics, **35**: 281–293.

Kish, L. 1957. Confidence intervals for clustered samples. Am. Soc. Rev. **22**: 154–165.

Kupper, L.L., and Haseman, J.K. 1978. The use of a correlated binomial model for the analysis of certain toxicological experiments. Biometrics, **34**: 69–76.

Nelder, J.A., and Mead, R. 1965. A simplex method for function minimization. Comput. J. **7**: 308–313.

Paul, S.R. 1982. Analysis of proportions of affected foetuses in teratological experiments. Biometrics, **38**: 361–370.

Rao, J.N.K., and Scott, A.J. 1992. A simple method for the analysis of clustered binary data. Biometrics, **48**: 577–585.

Tarone, R.E. 1979. Testing the goodness of fit of the binomial distribution. Biometrika, **66**: 585–590.

USDA Forest Service. 1988. A killer of dogwood: dogwood anthracnose. USDA For. Serv. South. Reg. Prot. Rep. R8-PR10.

USDA Forest Service. 1991. Results of the 1990 dogwood anthracnose impact assessment and pilot test in the southeastern United States. USDA For. Serv. South. Reg. Prot. Rep. R8-PR20.

Williams, D.A. 1975. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. Biometrics, **31**: 949–952.