ALTERNATIVES TO THE CLASSICAL FREQUENTIST CONFIDENCE INTERVAL FOR DESCRIBING ZERO-INFLATED LEAF DISEASE SEVERITY

Jhessica Letícia KIRCH[1]
Brena Geliane FERNEDA[2]
Fernando Henrique Silva GARCIA[3]
Sônia Maria De Stefano PIEDADE[1]
Idemauro Antonio Rodrigues de LARA[1]

- *ABSTRACT:* This paper presents the bootstrap percentile interval and the Bayesian credible interval as alternatives to the classical frequentist confidence interval for analysis of zero-inflated data. The indicated methods were applied to soybean downy mildew severity data obtained by stratified sampling in two municipalities in the state of São Paulo: Estiva Gerbi and Piracicaba. The amplitudes of the frequentist and bootstrap percentile confidence intervals were similar. For the Bayesian approach, the credible intervals of the posterior predictive distribution were considered using the zero-inflated beta distribution as likelihood. The credible intervals showed a wider range and included values in the upper bounds of the intervals greater than those observed in the data. We conclude that Bayesian inference is more complex, but allows incorporation of prior information regarding regional and seasonal aspects, contributing to better disease management in the field. When this information is not known, nonparametric bootstrap resampling is a simple alternative to construct intervals for zero-inflated data without assuming the distribution function.

- *KEYWORDS:* Stratified sample; Bayesian inference; *Perenospora manshurica*; Bootstrap sample.

1 Introduction

Brazil is one of the main producers and exporters of agricultural commodities in the world, standing out for its soy production. According to USDA (2020), world soybean production is led by Brazil (131 million tons), followed by the United States (112.2 million tons) and Argentina (53.5 million tons). These countries together account for about 80% of the world soybean production. It is estimated that the average productivity of the crop in Brazil is 3,542 kg ha^{-1} (CONAB, 2020), with variability between regions.

[1] Universidade de São Paulo - USP, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciências Exatas, 13418-900, Piracicaba, São Paulo, Brasil. E-mail: *jhessicakirch@gmail.com; soniamsp@usp.br; idemauro@usp.br*
[2] Universidade de São Paulo - USP, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Biossistemas, 13418-900, Piracicaba, São Paulo, Brasil. E-mail: *brenaferneda95@gmail.com*
[3] Universidade Federal do Amapá -UNIFAP, Departamento de Ciências Biológicas e Saúde, 68903-419, Macapá, AP, Brasil. E-mail: fernandogarcia@unifap.br

Fungal diseases occur in several soybean producing regions in Brazil (SILVA *et al.*, 2011) and worldwide (DUNLEAVY, 1987; LIM, 1989), leading to a reduction in productivity of up to 8% (PHILLIPS, 1999). Downy mildew is a fungal disease considered secondary in soybeans, caused by the pathogen *Perenospora manshurica*, which infects the unifoliate and trifoliate leaves. The favorable environmental conditions for the development of the pathogen in soybeans are air temperature ranging between 20 ºC and 22 ºC and a leaf wetness period of 12 hours (PICININI and FERNANDES, 2003).

The evolution of pathogen infection in the plant is determined preferentially by severity, which refers to the percentage of stunting or damage caused to the plant organ (leaf, stem, fruit). This determination factor, despite being more laborious compared to the evaluation of incidence, provides a better characterization of the evolution of the disease or the level of resistance of the plant to the pathogen, expressing more accurately the damage caused to the plant.

Disease severity can be measured visually, and for different assessors to follow the same rigor, diagrammatic scales and descriptive scales have been developed that contain a sequence of scores to be assigned according to the amount of damage to the plant organ. A challenge in modeling of severity data is zero inflation, occurring when there is no observation of the disease in the sampling unit. In this case, the data do not follow a normal distribution and usual methods can cause inaccurate parameter estimation and misleading inferences (MALDONADO *et al.*, 2016; MARTIN *et al.*, 2005).

The accuracy of estimating crop disease severity is dependent on the number of samples and the sampling scheme. Different sampling methods have been used in studies to assess foliar disease severity and incidence, including cluster sampling (MORALEJO *et al.*, 2019), systematic sampling (MICHEREFF *et al.*, 2008; ERTÜRK *et al.*, 2018; SONGTAO *et al.*, 2019), and stratified sampling (HARDWICK *et al.*, 2001; MANGENI *et al.*, 2020).

The goal of stratified sampling schemes is to reduce the overall sample size by controlling for sources of variability in the data (ABBOTT, 2010). It generally provides more accurate estimates of population characteristics from a small number of samples than simple random sampling (LIEB, 2020).

Confidence intervals are used to assess the reliability of estimates obtained from a random sample (SEVERIANO *et al.*, 2011; BRITO *et al.*, 2019). Confidence intervals can be determined by frequentist or computationally intensive methods. Frequentist inference is based on the assumption of the existence of a probabilistic model from which the random sample was drawn. If this model is not known, the inference may be compromised (FERREIRA, 2014).

Computationally intensive methods, such as nonparametric bootstrap resampling, empirically estimate the sampling distribution, and the only assumption required is that the sample be representative of the population (ZIENTEK and THOMPSON, 2007). Thus, they allow the construction of confidence intervals without having to assume a distribution function (SEVERIANO *et al.*, 2011) and can be an alternative to constructing confidence intervals for zero-inflated data, for which the distributional assumptions of classical models are generally not validated.

Bayesian inference can also be used as an alternative to interval estimation in the case of zero-inflated data, because it provides greater flexibility in modeling the data. Thus, Bayesian credible intervals can be considered as an alternative procedure for assessing the reliability of the mean severity estimation.

In this context, the present work focuses on the problem of constructing confidence and credible intervals for leaf fungal disease severity, presenting alternative methods for zero-inflated data and verifying the reliability of the estimated parameter of downy mildew severity in soybean crops in studies applying the stratified sampling scheme.

2 Methodology

2.1 Survey and sample collection

The study was conducted with soybean cultivar NEO660 IPRO, considered susceptible to downy mildew, in two experimental fields: in the municipalities of Estiva Gerbi (-22.206 ; -46.969) and Piracicaba (-22.698; -47.642), both in the state of São Paulo. During the experiment, Piracicaba and Estiva Gerbi registered average temperatures of 25.2 °C and 20.6 °C and accumulated precipitation of 900 mm and 1,114 mm between October and March. Soybeans were sown at a spacing of 0.5 m in 3×5 plots (6 rows, spaced at 0.5 m with a length of 5 m), representing a population of 320 thousand plants ha^{-1} in both locations.

The evaluations of downy mildew severity were performed at seven-day intervals from the first detection of the disease on the plant leaves (January 17, 2020) until the end of the crop cycle (February 21, 2020). Each evaluation was performed independently, with a new randomization performed weekly, so that although 6 weeks of evaluation were considered, the experimental units are independent in the evaluations. The severity of downy mildew was determined according to the diagrammatic scale described by Kowata *et al.* (2008), which was prepared according to logistic models so that severity levels compatible with human visual acuity were established.

The diagrammatic scale proposed by Kowata *et al.* (2008) to evaluate the severity of downy mildew in soybean caused by *Peronospora manshurica* consists of eight severity levels, 0.08%; 0.3%; 1.10%; 3.39%; 12.85%; 34.92%; 66.13% and 87.65%, which help in leaf evaluation and in determining the percentage of leaf area covered with symptoms.

Considering the difference in climate and altitude (60 m) between the municipalities, the severity evaluations were performed using the stratified sampling method in two strata (Estiva Gerbi and Piracicaba), and within each stratum the disease was evaluated by simple random sampling, with 60 samples per location, assuming values in percentage scale and admitting the non-observation of symptoms (0% severity), which occurred on average in 21% of the sample units, reaching 71% in the first week of evaluation in Piracicaba and 50% in the last week of evaluation in Estiva Gerbi.

2.2 Frequentist approach

The accuracy of the mean severity estimate can benefit from the construction of confidence intervals for the parameters. Since the estimate depends on a particular sample taken from the population, there is variability in the point estimates obtained from the samples relative to those obtained from the population (PINTO *et al.*, 2008).

Let y be the observed value of downy mildew severity, given in percent, and h be the stratum or location. In this notation, y_{ih} refers to the i-th observation of the h-th stratum, with $h = 1, 2$ and $i = 1,2, \ldots, n_h$, where $n_h = 60$ is the sample size of each stratum.

For the h-th stratum, a $100\gamma\%$ confidence interval for the mean considering simple random sampling across strata is defined according to Abbott (2010):

$$CI(\mu_h, 100\gamma\%) = \left(\bar{y}_h \mp z_{\frac{\alpha}{2}}\sqrt{\frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)}\right)$$

where n_h and N_h are, respectively, the sample and population sizes of stratum h and:

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$$

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$$

represents the sample mean and variance of stratum h.

An analytical expression for a $100\gamma\%$ confidence interval for the population mean in stratified sampling was described by Bolfarine and Bussab (2005):

$$CI(\mu, 100\gamma\%) = \left(\bar{y}_{st} \mp z_{\frac{\alpha}{2}}\sqrt{s_{\bar{y}_{st}}^2}\right)$$

where

$$\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{2} N_h \bar{y}_h$$

and

$$s_{\bar{y}_{st}}^2 = \frac{1}{N^2}\sum_{h=1}^{2} N_h^2 \frac{s_h^2}{n_h}$$

represent the self-weighted point estimators of the population mean and variance, which are unbiased. In this classical approach, confidence intervals were based on asymptotic normality.

Confidence intervals are often misinterpreted. A 95% confidence interval simply means that if the study is conducted multiple times (multiple sampling of the same population) and corresponding 95% confidence intervals have been constructed for the population mean, 95% of these intervals are expected to contain the true population mean (TAN and TAN, 2010).

2.3 Bootstrap Resampling

The bootstrap method uses resampling of the selected sample from the population to estimate the sampling distribution of a statistical estimator and obtain inferences about unknown population parameters (HANLEY and MACGIBBON, 2006; BERRAR, 2019).

In nonparametric bootstrap resampling, from a dataset containing n observations a number B of resampled datasets, called bootstrap samples, can be obtained. These resamples have size n, and the chosen estimator is applied to each one. Thus, let $y_1, y_2, \dots, y_n$ be an observed random sample and $\hat{\theta}$ be an estimate of θ. Then B resamples (or bootstrap samples) are selected and for each statistic of interest, denoted by $\hat{\theta}_i^*$, $i =$ 1,2, ..., and B is calculated. The point estimate will be given by:

$$\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

A confidence interval by the percentile method, obtained from the empirical distribution of bootstrap estimates, was defined by Efron and Tibshirani (1993):

$$\left[\hat{\theta}_{low}, \hat{\theta}_{upp}\right] = [\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}]$$

Thus, for $\gamma = 0.95$, we have:

$$\hat{\theta}_{low} = \hat{\theta}^{*(\alpha)} = 2.5\text{th percentile of the distribution of } \hat{\theta}^*$$
$$\hat{\theta}_{upp} = \hat{\theta}^{*(1-\alpha)} = 97.5\text{th percentile of the distribution of } \hat{\theta}^*$$

A direct extension of the usual bootstrap method for stratified samples is given below (RAO and WU, 1988):

1. Obtain a simple random sample with replacement $\{y_{hi}^*\}_{i=1}^{n_h}$ from the sample $\{y_{hi}\}_{i=1}^{n_h}$ in stratum h, independently for each stratum. Calculate:

$$\bar{y}_h^* = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}^* \qquad \text{and} \qquad \bar{y}^* = \sum_{h=1}^{2}\frac{N_h}{N}\bar{y}_h^*.$$

2. Calculate $\hat{\theta}^* = g(\bar{y}^*)$.
3. Independently repeat step 1 B times and calculate the corresponding estimates (step 2) $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \dots, \hat{\theta}^{*B}$.

Thus, a $100\gamma = 95\%$ confidence interval by the percentile method will be given as described by Sitter (1992) and Cusi (2007):

$$\left[\hat{\theta}_{low}, \hat{\theta}_{upp}\right] = [\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}]$$

The bootstrap resampling method used in the previous procedures is independent of the distribution, and is called the nonparametric bootstrap method.

2.4 Bayesian inference

In classical inference, the sample data y are considered random and the population parameters are considered fixed. In the Bayesian approach, the parameters are considered to follow a probability distribution, which expresses preliminary information about the process through an a prior distribution, called $\pi(\boldsymbol{\theta})$ (CONGDON, 2003). Thus, Bayesian analysis combines the information contained in the data with a prior information about the parameters (BOLSTAD, 2004). With the combination of the data distribution, $f(\boldsymbol{y}|\boldsymbol{\theta})$, and the prior distribution, $\pi(\boldsymbol{\theta})$, knowledge about the process is updated and summarized in a posterior distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, given by:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

In this context, considering the disease severity data, which assumes values on a percentage scale and may contain excess of zeros, it is necessary to use a probability distribution that incorporates the zero inflation and the range of variation. Models for zero-inflated data are combinations of probability distributions that separately model the occurrence of zeros and the other values in the domain of the variable of interest (MARTIN *et al.*, 2005). Thus, we considered in the model the zero-inflated beta distribution (BEZI), given by:

$$f_Y(y;\mu,\phi,\nu) = \begin{cases} \nu, & \text{if } y = 0 \\ (1-\nu)f(y;\mu,\phi), & \text{if } 0 < y < 1 \end{cases}$$

where ν is the mass probability of observing zeros ($0 < \nu < 1$) and $f(y;\mu,\phi)$ is the probability density function of the beta distribution with parameters μ and ϕ ($0 < \mu < 1$ and $\phi > 0$) proposed by Ferrari and Cribari-Neto (2004), which has probability density function given by:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} y^{(\mu\phi-1)}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1$$

where $\Gamma(\cdot)$ is the gamma function. Thus, the likelihood $L(\mu,\phi,\nu) = f(\boldsymbol{y}|\mu,\phi,\nu)$ considering a random sample with BEZI distribution will be given by:

$$L(\mu,\phi,\nu) = \prod_{i=1}^{n}\left[I_{(y_i=0)}\nu + I_{(0<y_i<1)}(1-\nu)f(y;\mu,\phi)\right]$$

$$L(\mu,\phi,\nu) = \prod_{i=1}^{n}\left[I_{(y_i=0)}\nu + I_{(0<y_i<1)}(1-\nu)\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} y^{(\mu\phi-1)}(1-y)^{(1-\mu)\phi-1}\right]$$

According to Congdon (2003), sometimes the existing knowledge is too imprecise to obtain an informative prior and therefore non-informative priors are used. Thus, the priors considered for the parameters were:

$$\mu \sim Uniform(0,1)$$
$$\phi \sim Uniform(0,+\infty)$$
$$\nu \sim Uniform(0,1)$$

The iterative Monte Carlo class algorithm using Markov Chain Monte Carlo (MCMC) was used for inference about the parameters μ, ϕ and ν. The convergence of the Markov chains was checked using the Gelman and Rubin (1992) criterion, which uses multiple repetitions of sequences to check whether stationarity has been reached within the second half of each sample, assuming that these sequences were simulated in parallel, each starting from a different starting point (NOGUEIRA *et al.*, 2004).

An advantage of the Bayesian approach is the more natural interpretation of random intervals for the parameters, called credible intervals (CONGDON, 2003). Credible intervals are given by the quantiles of the a posterior distribution, setting a credibility level γ, and are interpreted as a probability ($100\gamma\%$) of the true parameter belonging to the interval.

The predictive distribution of the mean severity was generated based on the a posterior estimates of the parameters given by the Bayesian approach for the posteriors by location/stratum. For analysis considering stratified sampling, the sample weight was considered similar to that considered for bootstrap stratified sampling.

2.5 Simulation Study

A simulation study has been conducted to compare the performance of confidence and credible intervals in simulated samples of the BEZI distribution, based on the posterior distribution of the parameters estimated in the Bayesian analysis. The simulation flowchart was as follows:

1. Random samples were generated of the BEZI distribution with sample size n =60 and parameters:
    a. μ ranging from 0.01 to 0.04.
    b. ϕ assuming values 50, 100 and 400.
    c. ν ranging from 0.1 to 0.9.
2. Frequentist and bootstrap confidence intervals and credible intervals were calculated for each sample and it was verified if the value of μ was covered in the estimation intervals.
3. Steps 1 and 2 were repeated 1000 times and the coverage probability was calculated.

3 Application

In Piracicaba, the disease severity was increasing until the fourth week of evaluation, in which the disease reached 10% damage to the leaves. However, from the fifth week on, there was a decrease in the severity of downy mildew until the end of the crop cycle. In Estiva Gerbi, the severity was relatively uniform throughout the crop cycle, with peaks of 5% and 10% in the third and sixth week of evaluation, respectively. In both municipalities analyzed, the severity values were on average less than 3%. These results show that the disease was present in the plants (samples), but not in the entire population. In Estiva Gerbi, we observed values of mildew severity on soybean from 0.879±0.98 to 1.466±1.01, while in Piracicaba these values ranged from 0.315±0.82 to 4.035±2.53.

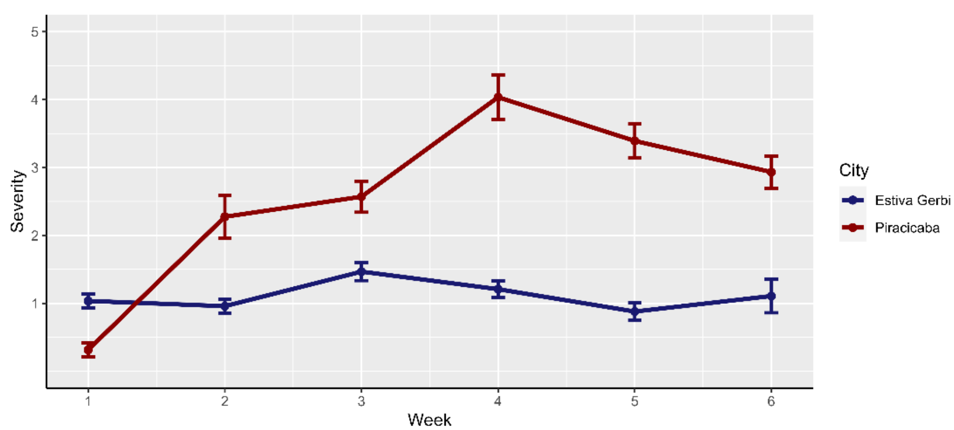

Figure 1 - Mean and standard error of the average severity of downy mildew in soybean in the municipalities of Piracicaba and Estiva Gerbi in 2020.

The average severity of downy mildew differed between Piracicaba and Estiva Gerbi throughout the evaluations. Only in the first week of evaluation did Estiva Gerbi show higher average severity than Piracicaba (Figure 1). After the second week of evaluation, the average severity of downy mildew was higher in Piracicaba than in Estiva Gerbi.

The amplitudes of the confidence intervals were similar for the frequentist and bootstrap methods (Figure 2), but the assumption of normal or approximately normal distribution of the estimator of the parameter of interest was rejected by the Shapiro-Wilk normality test. The confidence intervals in Piracicaba for stratified sampling showed a similar trend, but in Estiva Gerbi this trend was not present.

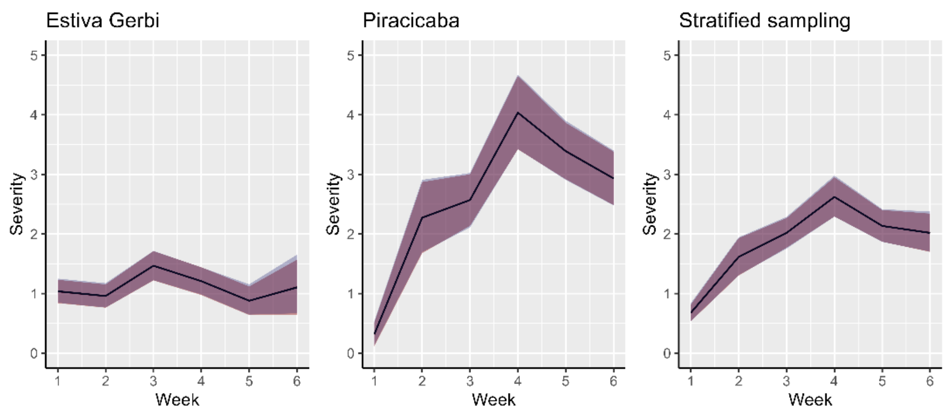

Figure 2 - Frequentist and bootstrap confidence intervals for the average severity of downy mildew on soybean in the municipalities of Piracicaba and Estiva Gerbi in the stratified sampling in 2020.

For the Bayesian approach (Table 1), the credible intervals of the posterior predictive distribution were considered. The Gelman and Rubin (1992) criterion was used to verify the convergence of the chains, which was achieved considering 11,000 iterations. The values of μ, which represents the mean of the zero-inflated beta distribution, did not vary greatly, but the values of ϕ, which is a precision parameter, varied between 59.53 and 468.51 in Estiva Gerbi and between 39.05 and 99.58 in Piracicaba.

Table 1 - Posterior estimates for the parameters of the BEZI distribution in the study of the average mildew severity in soybean in the municipalities of Piracicaba and Estiva Gerbi in 2020, with μ representing the mean of the distribution, ϕ the precision parameter and ν modeling the zero inflation

	Estiva Gerbi			Piracicaba		
	μ	ϕ	ν	μ	ϕ	ν
Week 1	0.01	468.51	0.29	0.01	68.88	0.71
Week 2	0.01	263.09	0.19	0.03	39.05	0.14
Week 3	0.02	137.88	0.03	0.03	58.12	0.07
Week 4	0.01	184.06	0.19	0.04	57.17	0.02
Week 5	0.01	147.34	0.37	0.04	99.58	0.05
Week 6	0.02	59.53	0.50	0.03	88.13	0.06

The posterior mean of ν, which models the zero inflation, represented approximately the proportion of zeros observed in the sample, being more accurate the higher the percentage of zeros. In the evaluations with the lowest proportions of zeros observed in the sample, the posterior mean of ν was slightly overestimated, but still belonged to the 95% credible interval (Figure 3).

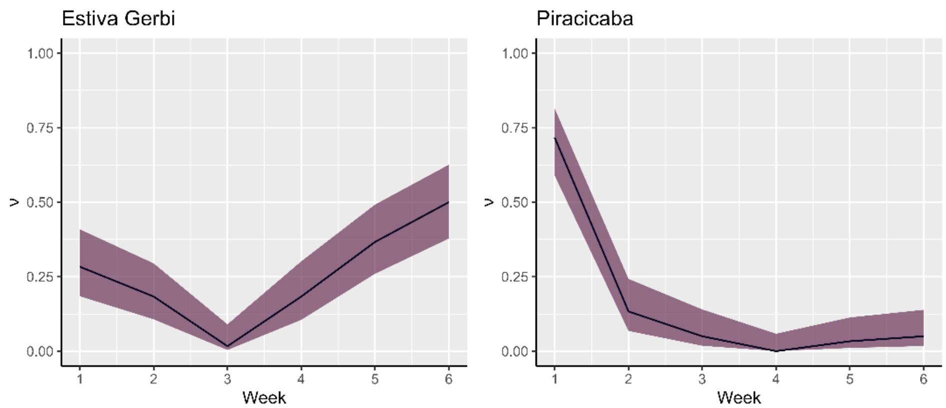

Figure 3 - Credible intervals for the parameter ν that models the zero inflation in the study of the average mildew severity in soybean in the municipalities of Piracicaba and Estiva Gerbi in 2020. The centerline is the proportion of zeros observed in the sample.

The credible interval was wider in Piracicaba (Figure 4) than in Estiva Gerbi and in the stratified sampling interval. In Estiva Gerbi, the credible interval presented the greatest amplitude in the sixth week of evaluation, while in Piracicaba, the greatest amplitude occurred in the fourth week of evaluation (Figure 4).

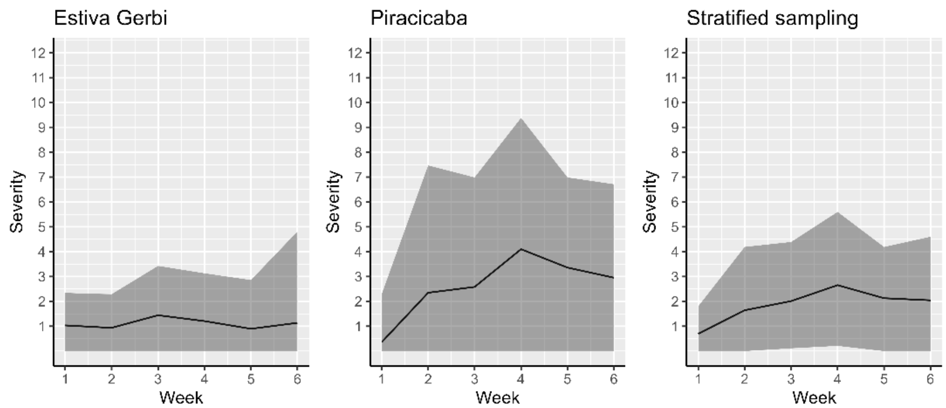

Figure 4 - Credible intervals for the average severity of downy mildew in soybean in the municipalities of Piracicaba, Estiva Gerbi and in the stratified sampling in 2020.

The credible intervals included the value zero in all evaluations with the exception of the fourth week of evaluation in the stratified sampling, and the upper limits of the intervals presented values higher than those observed in the data. However, the average severities presented values similar to those observed in the frequentist and bootstrap percentile intervals.

Using the posterior estimates for the parameters of the BEZI distribution, a simulation study was conducted considering $n = 60$, μ ranging from 0.01 to 0.04, ϕ assuming values 50, 100 and 400 and ν ranging from 0.1 to 0.9. It was observed that the coverage probability was similar for the frequentist and bootstrap confidence intervals, reaching the value zero around $\nu = 0.5$ in all scenarios. The credible intervals, on the contrary, maintained a coverage probability close to 100% until $\nu = 0.8$ in all scenarios (Figure 5). Our results demonstrate that the frequentist and bootstrap confidence intervals and the credible intervals differ in coverage probability, being mainly affected by the parameter ν that models the zero-inflation in the sample.

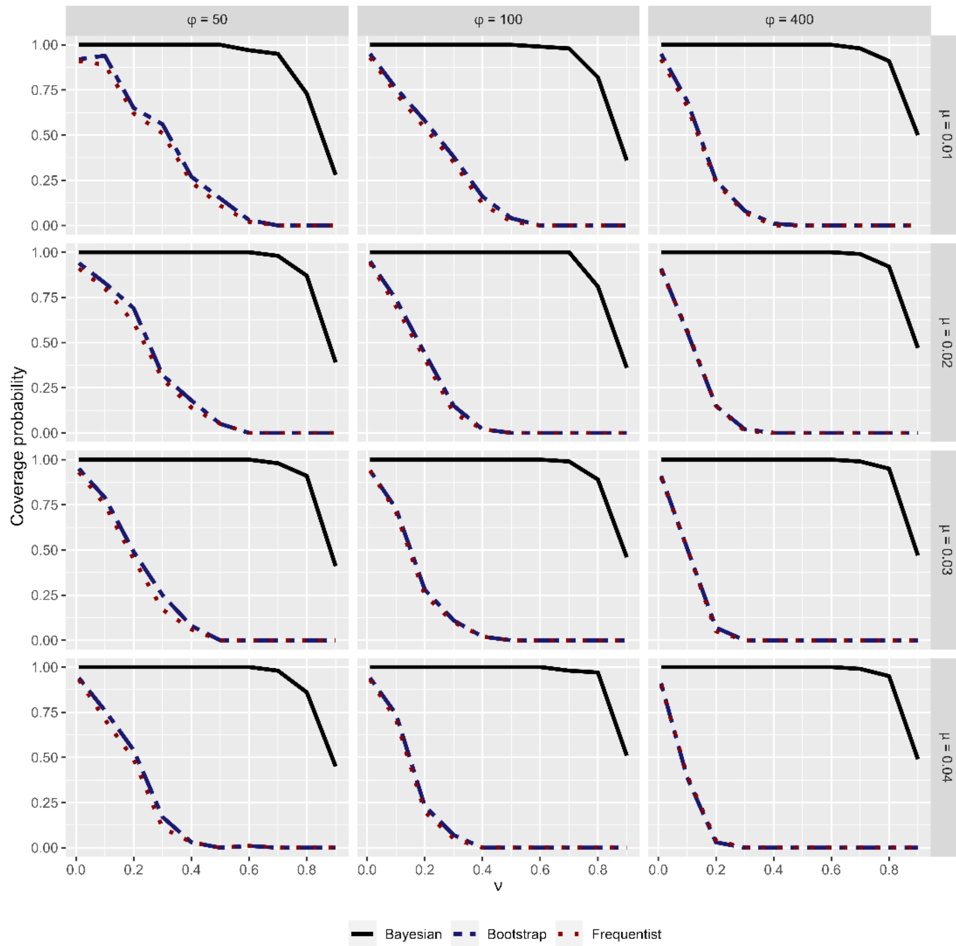

Figure 5 - Coverage probability of frequentist and bootstrap confidence intervals and credible intervals as a function of the ν parameter of the BEZI distribution with μ assuming values ranging from 0.01 to 0.04, ϕ assuming values 50, 100 and 400 and ν ranging from 0.1 to 0.9 for a sample size n =60.

4 Discussion and concluding remarks

In this study, we observed severity values of downy mildew in soybean ranging from 0 to 10%, which are considered low compared to other studies, such as that of Kowata *et al.* (2008), who reported severity levels of downy mildew ranging from 0.08% to 87.65%, reaching 90% in the most advanced stage of the disease.

The difference in mildew severity observed between the municipalities of Estiva Gerbi and Piracicaba in the cycle may be associated with the difference in environmental conditions between the municipalities. In the first week of evaluation, Estiva Gerbi exhibited higher severity of downy mildew than Piracicaba, probably due to higher rainfall, since the pathogen needs a leaf wetness period of 12 hours and a temperature between 20 and 22 °C to infect the plant (PICININI and FERNANDES, 2003). However, the increase in downy mildew severity in soybean in Piracicaba may be associated with lower plant resistance caused due to greater water and nutrient limitation. Some nutrients, such as nitrate and sulfate, are present in the soil solution, being absorbed mostly by mass flow. Therefore, the lower rainfall may reduce nutrient uptake by soybean, which may consequently reduce plant resistance to the pathogen.

In foliar diseases, the percentage of tissue area affected (severity) allows for a better evaluation of the impact of the disease on the plant than the incidence. Diagrammatic scales help in this quantification, although attention must be paid to the limitations of the acuity of human vision to observe the symptoms on the plant (AMORIM, 1995). However, Silva *et al.* (2009) found that even with the aid of a diagrammatic scale, experienced assessors did not accurately estimate the severity, casting doubt on the human ability to distinguish some types of leaf lesions when very small, and to condense all visual stimuli into a numerical estimate. This contrasts with the accuracy obtained by using software to quantify the actual severity, which possibly identifies injured tissue that is impractical within the range of human vision (SILVA *et al.*, 2009).

Thus, the reliability of the severity estimates obtained was evaluated using confidence intervals and credible intervals. From these intervals, it is possible to identify upper and lower limits and evaluate the accuracy of the estimate, which is necessary for appropriate decision making in disease control.

The ranges of the frequentist and bootstrap confidence intervals for downy mildew severity in soybean were similar. Other authors have also observed such similarities for the frequentist and bootstrap confidence intervals in simple random sampling (PEREIRA *et al.*, 2000) and stratified sampling (CUSI, 2007).

Despite the similarity of the results, the frequentist confidence interval is only valid under the assumption of normal or approximately normal distribution of the parameter estimator of interest, which in this study was rejected by the Shapiro-Wilk normality test. Carpenter and Bithell (2000) suggest that bootstrap confidence intervals should be used whenever there is a reason to doubt the assumptions related to classical confidence intervals. Carrasco *et al.* (2012), who used the parametric and nonparametric bootstrap approach to obtain confidence intervals considering data with zero inflation, concluded that resampling is an alternative procedure that makes it possible to obtain adequate confidence intervals.

Another alternative to analyze zero-inflated data is Bayesian inference, which allows incorporating in the analysis a prior information regarding regional and seasonal aspects (BRIGHENTI *et al.*, 2011; GARTHWAITE *et al.*, 1995). This is especially useful for leaf fungal disease severity data, since disease development is associated with these aspects. Thus, obtaining Bayesian credible intervals can help improve the accuracy of estimating the disease severity in the plant when considering a prior information or information from previous experiments to update this information with new data, contributing to the better management of the disease in the field.

In this study, severity values equal to zero were observed, so the zero-inflated beta distribution was considered in Bayesian inference for the data. The posterior parameters (Table 1) show mean severity close to zero, ranging from 1 to 4%. The value of ϕ (precision parameter), was higher in Estiva Gerbi, causing a smaller amplitude in its confidence interval (Figure 4). Regarding the parameter that models the zero inflation (ν), there was a decreasing pattern in Piracicaba throughout the evaluations, showing that the chance of observing leaves without the disease decreased throughout the crop cycle. Considering the possibility of observing the zero value is important for the management of the disease, since unnecessary treatment, especially in the case of chemical control, can lead to ecological and economic losses.

This article is an introduction to the problems related to the construction of classical frequentist confidence intervals for disease severity data, especially for zero-inflated data. The measurement and quantification of this parameter is important to prevent the spread of disease and to make decisions about crop protection measures, considering the costs of disease control and the resulting economic and productive losses.

It should be considered that bootstrap resampling and Bayesian inference use simulations, but they are more flexible for data modeling, and may be useful for obtaining intervals for zero-inflated data. We conclude that Bayesian inference is a more complex method, but allows incorporation of prior information regarding regional and seasonal aspects, contributing to better disease management in the field. When this information is not known, nonparametric bootstrap resampling is a simple alternative to construct intervals for zero-inflated data without the need to know the analytical expressions.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

KIRCH, J. L.; FERNEDA, B. G.; GARCIA, F. H. S.; PIEDADE, S. M. S.; LARA, I. A. R. Alternativas ao intervalo de confiança clássico frequentista para descrição da severidade de doença foliar com inflação de zeros. *Braz. J. Biom*. Lavras, v.40, n.2, p.181-197, 2022.

▪ *RESUMO: Este trabalho apresenta o intervalo percentil bootstrap e o intervalo de credibilidade bayesiano como alternativas ao intervalo de confiança clássico frequentista para análise de dados com inflação de zero. As metodologias indicadas foram aplicadas a dados de severidade do míldio na soja, obtidas por amostragem estratificada em duas cidades do estado de São Paulo: Estiva Gerbi e Piracicaba. As amplitudes dos intervalos de confiança frequentista e percentil bootstrap foram aproximadamente iguais. Para a abordagem bayesiana foram considerados os intervalos de credibilidade da distribuição preditiva a posteriori utilizando a distribuição beta inflacionada de zeros como verossimilhança. Os intervalos de credibilidade apresentaram uma maior amplitude e incluíram nos limites superiores dos intervalos valores acima dos observados nos dados. Concluiu-se que a inferência bayesiana apresenta uma metodologia mais complexa, porém permite incorporar informação a priori referente a aspectos regionais e sazonais, contribuindo* *para o melhor manejo da doença no campo. Quando não se conhece essas informações, a reamostragem bootstrap não paramétrica é uma alternativa simples para construção de intervalos para dados inflacionados de zeros sem que seja necessário assumir uma função de distribuição para a mesma.*

- *PALAVRAS-CHAVE: Amostragem estratificada; Inferência bayesiana; Perenospora manshurica; Reamostragem Boostratap.*

References

ABBOTT, A. M. Statistical sampling methods for soils monitoring. In: PAGE-DUMROESE, D.; NEARY, D.; TRETTIN, C. (Eds.). Scientific background for soil monitoring on National Forests and Rangelands: workshop proceedings. *USDA Forest Service Proceedings*, p.109–120, 2010.

AMORIM, L. Avaliação de doenças. In: BERGAMIN FILHO, A. E. A. (Ed.) Manual de fitopatologia. São Paulo: *Agronômica Ceres*, p.647–671, 1995.

BERRAR, D. Introduction to the non-parametric bootstrap.In: RANGANATHAN, S.; GRIBSKOV, M.; NAKAI, K.; SCHONBACH, C. (Eds.) *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press, p.766-773, 2019.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de amostragem*. ABE - Projeto Fisher. Edgard Blucher, 2005.

BOLSTAD, W. M. *Introduction to Bayesian statistics*. New Jersey: John Wiley & Sons, Inc., 2004.

BRIGHENTI, C. R. G.; RESENDE, M.; BRIGHENTI, D. M. Estimação sequencial Bayesiana aplicada à proporção de infestação de psilídeos em alecrim do campo. *Revista Brasileira de Biometria*, v.29, n.2, p.342–354, 2011.

BRITO, O.; ANDRADE JÚNIOR, V. C. D.; AZEVEDO, A.; DONATO, L.; SILVA, L.; FERREIRA, M. Study of repeatability and phenotypical stabilization in kale using frequentist, bayesian and bootstrap resampling approaches. *Acta Scientiarum. Agronomy*, v.41, p.42606, 2019.

CARPENTER, J.; BITHELL, J. Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, v.19, p.1141-1164, 2000.

CARRASCO, C. G.; TUTIA, M. H.; NAKANO, E. Y. Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. *TEMA (São Carlos)*, v.13, p.247–255, 2012.

CONAB. Companhia Nacional de Abastecimento . *Acomp.safra brasileira de grãos, v.8 – safra 2020/21, n.1 - primeiro levantamento*. Brasília, p.1–77, 2020.

CONGDON, P.*Applied Bayesian modelling*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2003.

CUSI, M. L. A. Estimación de la distribución estadística de la tasa global de fecundidad. *Papeles de población*, v.13, n.54, p.87-113, 2007.

DUNLEAVY, J. M. Yield reduction in soybean caused by downy mildew. *Plant Disease*, v.71, n.12, p.1112-1114, 1987.

EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap.Number 57 in Monographs on Statistics and Applied Probability*. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1993.

ERTÜRK, H.; KARAKAYA, A.; ÇELIK OGUZ, A. Leaf diseases occurring on barley plants in bala district of ankara province, *turkey.e-Journal of New World Sciences Academy*, v. 13, p.204–207, 2018.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of applied statistics*, v.31, n.7, p.799–815, 2004.

FERREIRA, D. F. Sisvar: a guide for its bootstrap procedures in multiple comparisons. *Ciência e Agrotecnologia*, v.38, p.109–112, 2014.

GARTHWAITE, P.H.; JOLLIFFE, I.; BYRON, J. *Statistical inference*. 2. ed. London: Prentice Hall, 1995.

GELMAN, A.; RUBIN, D. Inference from iterative simulation using multiple sequences. *Statistical Science*, v.7, p.457–511, 1992.

HANLEY, J. A.; MACGIBBON, B. Creating non-parametric bootstrap samples using poisson frequencies. *Computer Methods and Programs in Biomedicine*, v. 83, n.1, p.57–62, 2006.

HARDWICK, N.; JONES, D.; SLOUGH, J. Factors affecting diseases of winter wheat in england and wales, 1989–98. *Plant Pathology*, v.50, p.453–462, 2001.

KOWATA, L.; MAY-DE MIO, L.; DALLA PRIA, M.; SANTOS, H. Escala diagramática para avaliar severidade de míldio na soja. *Scientia agraria*, v.9, n.1, p.105–110, 2008.

LIEB, M. At the interface between domain knowledge and statistical sampling theory: Conditional distribution based sampling for environmental survey (codibas). *Catena*, v.187, p.1–10, 2020.

LIM, S. M. Inheritance of resistance to *peronospora manshurica* race 2 and race 33 in soybean. *Phytopathology*, v.79, p.877–879, 1989.

MALDONADO, A.; AGUILERA, P.; SALMERÓN, A. Modeling zero-inflated explanatory variables in hybrid bayesian network classifiers for species occurrence prediction. *Environmental Modelling & Software*, v.82, p.31-43, 2016.

MANGENI, B. C.; WERE, H. K.; NDONG'A, M.; MUKOYE, B. Incidence and severity of bean common mosaic disease and resistance of popular bean cultivars to the disease in western kenya. *Journal of Phytopathology*, v.168, p.1–15, 2020.

MARTIN, T.; WINTLE, B.; RHODES, J.; KUHNERT, P.; FIELD, S.; LOWCHOY, S.;TYRE, A.; POSSINGHAM, H. Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, v.8, p.1235-1246, 2005.

MICHEREFF, S. J.; NORONHA, M. D. A.; MAFFIA, L. A. Tamanho de amostras para avaliação da severidade da queima das folhas do inhame. *Summa Phytopathologica*, v.34, n.2, p.189-191, 2008.

MORALEJO, E.; BORRAS, D.; GOMILA, M.; MONTESINOS, M.; ADROVER, F.; JUAN, A.; NIETO, A.; OLMO, D.; SEGUI, G.; LANDA, B. Insights into the epidemiology of pierce's disease in vineyards of mallorca, spain. *Plant Pathology*, v.68, 2019.

NOGUEIRA, D. A.; SÁFADI, T.; FERREIRA, D. F. Avaliação de critérios de convergência univariados para o método de Monte Carlo via cadeias de Markov. *Revista Brasileira de Estatística*, v.65, n.224, p.59–88, 2004.

PEREIRA, J. E.; SILVA, J. F. V.; DIAS, W. P.; SOUZA, G. S. Intervalo de confiança "bootstrap" como ferramenta para classificar raçaas do nematoide de cisto da soja. *Pesquisa Agropecuária Brasileira*, v.35, p.271–275, 2000.

PHILLIPS, D. V. Downy mildew. In: HARTMAN, G.; SINCLAIR, J.; J.C., R. (Eds.) *Compendium of soybean diseases*. 4.ed. St. Paul: APS Press, 1999.

PICININI, E. C.; FERNANDES, J. M. *Doenças de soja: Diagnose, epidemiologia e controle*. 3.ed. Passo Fundo, RS: Embrapa, 2003.

PINTO, F.; MELO-CRISTINO, J.; RAMIREZ, M. A confidence interval for the wallace coefficient of concordance and its application to microbial typing methods. *PloS one*, v.3, p.e3696, 2008.

RAO, J. N. K.; WU, C. F. J. Resampling inference with complex survey data. *Journal of the American Statistical Association*, v.83, n.401, p.231–241, 1988.

SEVERIANO, A.; CARRIÇO, J.; ROBINSON, D.; RAMIREZ, M.; PINTO, F. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures. *PloS one*, v.6, p.e19539, 2011.

SILVA, O.; SANTOS, H.; DALLA PRIA, M.; MAY-DE MIO, L. Potassium phosphite for control of downy mildew of soybean. *Crop Protection*, v.30, n.6, p.598-604, 2011.

SILVA, S. X. B.; LARANJEIRA, F. F.; SOARES, A. C. F.; MICHEREFF, S. M. Amostragem, caracterização de sintomas e escala diagramática da mancha graxa dos citros (*mycosphaerella citri*) no recôncavo baiano. *Ciência Rural*, v.38, p.896–899, 2009.

SITTER, R. R. A resampling procedure for complex survey data. *Journal of the American Statistical Association*, v.87, n.419, p.755–765, 1992.

SONGTAO, B.; TIAN, M.; CHANG, Q. Estimating the severity of apple mosaic disease with hyperspectral images. *International Journal of Agricultural and Biological Engineering*, v.12, p.148–153, 2019.

TAN, S.; TAN, S. The correct interpretation of confidence intervals. *Proceedings of Singapore Healthcare*, v.19, p.276–278, 2010.

USDA. Departamento de agricultura dos Estados Unidos. *Relatórios USDA*. 2020.

ZIENTEK, L. R.; THOMPSON, B. Applying the bootstrap to the multivariate case: Bootstrap component/factor analysis. *Behavior Research Method*, v.39, n.2, p.318–325, 2007.

Received in 23.06.2021

Approved after revised in 25.09.2021