SCIENTIFIC NOTE

NON-PARAMETRIC STATISTICAL METHODS AND DATA TRANSFORMATIONS IN AGRICULTURAL PEST POPULATION STUDIES

Alcides Cabrera Campos[1*], Caridad W. Guerra Bustillo[2], Magaly Herrera Villafranca[3], and Moraima Suris Campos[4]

Analyzing data from agricultural pest populations regularly detects that they do not fulfill the theoretical requirements to implement classical ANOVA. Box-Cox transformations and nonparametric statistical methods are commonly used as alternatives to solve this problem. In this paper, we describe the results of applying these techniques to data from *Thrips palmi* Karny sampled in potato (*Solanum tuberosum* L.) plantations. The χ^2 test was used for the goodness-of-fit of negative binomial distribution and as a test of independence to investigate the relationship between plant strata and insect stages. Seven data transformations were also applied to meet the requirements of classical ANOVA, which failed to eliminate the relationship between mean and variance. Given this negative result, comparisons between insect population densities were made using the nonparametric Kruskal-Wallis ANOVA test. Results from this analysis allowed selecting the insect larval stage and plant middle stratum as keys to design pest sampling plans.

Key words: Kruskal-Wallis test, negative binomial distribution, Box-Cox transformations, *Thrips palmi*, *Solanum tuberosum*.

Statistical management of data in population studies of agricultural pests is one of the greatest challenges for researchers dedicated to plant protection; this occurs when they are faced with designing experiments, data analysis, and drawing conclusions. The parametric statistical methods are widespread and well-known and are the most used in these studies. Nevertheless, many researchers do not know that they are subjected to fulfilling theoretical assumptions, such as normality, variance homogeneity, and no correlation between errors. If these conditions are not fulfilled, the statistical analysis of the results can be invalidated (De Calzadilla *et al*., 2002; Santos *et al*., 2005).

The lack of normality of errors is of little importance in Fisher's F-test from ANOVA because it is a robust technique in the presence of deviations of this assumption. However, it may affect variance homogeneity mainly when there is a great difference in the number of observations in the groups or treatments. This heterogeneity is usually accompanied by non-normal variables, so it is recommended that transformations be applied to stabilize variances and normalize responses (Box and Cox, 1964; Peña, 1994; Font *et al*., 2007).

Menchaca (1973) considered the parametric family of transformations of Y in $Y_i^{(\lambda)}$, where λ defines a particular transformation, and it is assumed that for any unknown λ, the transformed observations $Y_i^{(\lambda)}$ ($i = 1, 2, \ldots, n$) fulfill the basic hypothesis, these elements are provided by Box and Cox (1964). Transformations are performed to search for a new scale in the analyzed variables in such a way that the errors are approximately normally distributed and have homogeneous variances (Eisenhart, 1947; Steel and Torrie, 1992).

Data of agricultural pest populations do not generally fulfill the basic assumptions to apply parametric statistical methods because they are essentially discrete. This explains why the variables are not generally adjusted to the normal distribution but to binomial, negative binomial, and Poisson discrete probabilistic distributions according to the spatial pattern of the individuals in their habitat (Sokal and Rohlf, 1995).

Nonparametric statistical methods are necessary as an alternative in these cases because they do not depend on data distribution, can be used for small samples, and are generally faster and simpler to apply (Siegel and Castellan, 1995; Gómez *et al*., 2003; Santos *et al*., 2005). If it is

[1]Universidad de las Ciencias Informáticas, Autopista Novia del Mediodía, km 2½, Torrens, Boyeros, La Habana, Cuba. [*]Corresponding author (alcides@uci.cu).
[2]Centro Universitario Municipal de Güines, Calle 86, N° 7312, entre 73 y 77, Güines, Mayabeque, Cuba.
[3]Instituto de Ciencia Animal, Apartado Postal 24, San José de las Lajas, Mayabeque, Cuba.
[4]Centro Nacional de Sanidad Agropecuaria, Autopista Nacional y Carretera de Tapaste, San José de las Lajas, Apartado Postal 10, Mayabeque, Cuba.
Received: 8 December 2011.
Accepted: 23 May 2012.

also considered that there is a relative lack of knowledge and feasibility about these methods, it is convenient to deal with some experiences obtained in agricultural pest population studies.

The objective of this study was to select the insect stage and the plant stratum of greater predominance as key elements to design sampling plans using data from agricultural pest populations with a negative binomial distribution, and according to the application of non-parametric statistical procedures and seven expressions to transform data.

MATERIALS AND METHODS

Data came from *Thrips palmi* Karny (Thysanoptera: Thripidae) populations sampled in potato (*Solanum tuberosum* L.) plantations established in production conditions during three seasons (winters of 1998-2000) in areas of Güira de Melena Municipality (22°47′26″ N, 82°30′19″ W), Artemisa Province, Cuba. Data from all three seasons were combined in a single set and the samples considered as coming from the same population were tested for homogeneity by Taylor's Power Law (TPL) regression slopes and intercepts and by an analysis of covariance to compare regression lines (Cabrera *et al*., 2008).

Crop areas sown with the Fregat irrigation system in circular form (43.2 ha) consisted of four quadrants. One of them was selected as an experimental area with five plots of 20 × 20 m. Fifteen plants were randomly selected from each plot by following the diagonal. The direction was changed in each sampling. Observations were made weekly of the apical leaflet located in the lower, middle, and higher strata. Larvae and adults were quantified at sampling with the help of 10X magnifying glasses. Ten samplings were carried out in each season.

The χ^2 test

It was applied to the variants of goodness-of-fit and of independence, respectively, to determine: a) The fit between the observed frequencies and those calculated by the negative binomial distribution to estimate its parameter k; and b) the relationship between plant strata and insect stages.

Data transformation

With the adjustment of the negative binomial distribution, data were transformed through the following expressions: $\sqrt{X+1}$, $\sqrt{X+0.5}$, $\sqrt{X+0.375}$, log (X+1), log (X+(k/2)), log [log (X+2)] and $X^{1-(b/2)}$

The values of k and b correspond to the parameter of the negative binomial distribution and the regression coefficient of Taylor's Power Law (TPL) as calculated by Cabrera (2002). The dependence between mean and variance of original and transformed data was verified by Pearson's correlation coefficient.

Nonparametric Kruskal-Wallis ANOVA test

The two fundamental advantages of this analysis, as compared with Fisher's statistical *F* from ANOVA, are: a) It does not require establishing assumptions, such as normality and homocedasticity on the original populations; and b) it allows working with ordinal data.

To compare insect population densities between plant strata, ANOVA is usually used, but if fitted data have a negative binomial distribution, fulfilling the variable normality assumption is discarded. To determine the adequate transformation to satisfy ANOVA assumptions, seven expressions were used; when assumptions were not fulfilled, the nonparametric Kruskal-Wallis ANOVA test was applied. The test of nonparametric multiple comparisons was used when necessary according to Conover (1999), who proposes using the usual parametric procedure, Fisher's least significant difference, which is computed on ranks instead of data.

Data were processed with Statgraphics Centurion XV statistical software (Stat Point Technologies, Warrenton, Virginia, USA) and InfoStat version 2008 (Di Rienzo *et al*., 2008).

RESULTS AND DISCUSSION

The following were the results of the χ^2 test to determine goodness-of-fit of the negative binomial distribution:

- There was a good fit in seven samplings for the larval stage of *T. palmi* with values of χ^2 between 3.52 and 5.39 and likelihood values higher than 0.05.
- It was not possible to apply this test in the first two samplings because densities were very low (0.0207 and 0.0593, respectively) because of short crop establishment. This reduced the number of classes, which was not greater than three; at least four classes were needed to fit this distribution.
- In *T. palmi* adult populations, frequencies were adequately fitted to the negative binomial distribution in five samplings. Those with no good fit obtained high values of χ^2 and a small number of classes, which also diminished the corresponding degrees of freedom.

The fit to this probabilistic model allows estimating the k parameter as an aggregation index in the expression log (X + (k/2)). This is used to transform data, calculate sample sizes, and design sequential sampling plans (Southwood, 1995).

For the χ^2 test of independence, the outcome showed there was a dependence between plant strata and insect stages (χ^2 = 473.83, d.f. = 2 and P ≤ 0.0001). Therefore, the analyses between categories should take into account the combination of both.

Santos *et al*. (2005) and Tinoco (2008) explained and exemplified the application of this statistical test to verify independence in the occurrence of events. Furthermore, Fernández *et al*. (2005) used it to find the relationship between the lower, middle, and higher strata of avocado

(*Persea americana* L.) plants and the stage of *Pseudacysta perseae* Heidemann nymphs and adults. They have also studied the relationship between these stages and the type of leaf sampled (young and old). Liscano and Domínguez (2005) applied this procedure to study the distribution of immature insects of the defoliating lepidopteron *Antichloris viridis* Druce in Venezuelan banana (*Musa paradisiaca* L.) plantations.

Seven data transformations were tested based on the fit of negative binomial distribution and variance heterogeneity inside the aggregate spatial pattern of this insect species and of the close relationship between the variance and the mean from the fit to the TPL (Cabrera, 2002). The results for larvae (Table 1) revealed that none of the expressions used eliminated dependence between mean and variance with coefficients of correlation (r) very close to one and statistically significant ($P \leq 0.0001$).

Nevertheless, for *T. palmi* in mango (*Mangifera indica* L.), Verghese *et al*. (1988) found that some of the transformations indeed eliminated the dependence between mean and variance, as well as stabilizing the latter. Among the transformations that did not solve the problem were $\sqrt{X}+1$ and $X^{1-(b/2)}$.

This result was similar to the one obtained in this study. It also agreed with Costello and Daane (1997), who reported this when trying to stabilize the variance and fulfill the ANOVA requirements in data from populations of different spider species in grape (*Vitis vinifera* L.) plantations. This deserves special attention because the first transformation is one of those most applied to data from insect populations and sometimes without verifying a priori and a posteriori the fulfillment of the assumptions. Other authors such as Miranda *et al*. (2004), Liscano and Domínguez (2005), and Font *et al*. (2007) have reached similar results when using square root and logarithmic type transformations.

De Calzadilla *et al*. (2002) noted that in data from trials conducted with a completely randomized design and random blocks using 78 discrete variables transformed by $\sqrt{X}$ or $\sqrt{X}+0.375$, it was observed that the transformation did not fulfill the theoretical assumptions of the models in 46.2% of the cases analyzed. In 42.3% of the cases, the transformation was improperly applied and assumptions were fulfilled in only 11.5%.

The lack of fulfillment of the assumptions for parametric ANOVA and dependence between plant strata and insect stages led to applying the nonparametric Kruskal-Wallis ANOVA test for the previous data combination of both categories. Therefore, the comparison of population densities was performed for the six resulting combinations (Table 2) by searching for the predominance of the insect stage and the site of higher concentration in plant foliage, which are essential elements to the sampling plans.

Larvae of the middle stratum were predominant (Table 2). This occurrence is understandable considering that individuals have better life conditions in this plant stratum because they are less exposed to the action of natural enemies and sun radiation. Besides, there are leaves in this level that, given their age, could be more adequate to guarantee food for the insect in larval stages (Suris and Plana, 2001; Tobing, 2007; Varga *et al*., 2010).

Table 2. Comparison of combinations between plant strata and *Thrips palmi* stages.

Combinations	Sum	Mean	Standard deviation	Mean ranges
Lower-Larvae	1869	1.08	3.19	5517.68b
Middle-Larvae	3097	1.79	4.91	5727.65a
Higher-Larvae	1779	1.03	3.17	5151.65c
Lower-Adults	255	0.15	0.52	4410.08e
Middle-Adults	611	0.35	0.96	4906.40d
Higher-Adults	964	0.56	1.17	5339.52bc

Statistics: H = 377.115, P = 0.0000, n = 1725.
Mean ranges with the same letter do not differ statistically according to the nonparametric multiple range test ($P < 0.05$).

CONCLUSIONS

The results obtained from applying statistical-methodological criteria and combining non-parametric statistical methods with data transformations allowed selecting the insect larval stage and plant middle stratum as the key elements to design sampling plans for this pest.

It was concluded that criteria provide alternative methods to improve conduction, analysis, and interpretation of research studies in agricultural pest populations, as well as informing decision-making in this type of work.

Table 1. Results of larvae data transformations to fulfill ANOVA requirements.

	X		T1		T2		T3		T4		T5		T6		T7	
Sampling	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
1	0.02	0.02	1.01	0.00	0.72	0.01	0.62	0.01	0.01	0.01	-2.52	0.14	-0.36	0.00	0.02	0.02
2	0.06	0.14	1.02	0.02	0.73	0.02	0.64	0.02	0.03	0.03	-2.47	0.32	-0.35	0.01	0.04	0.04
3	0.18	0.44	1.06	0.05	0.78	0.06	0.69	0.07	0.10	0.09	-2.26	0.87	-0.31	0.04	0.12	0.12
4	0.90	5.48	1.26	0.31	1.01	0.38	0.93	0.41	0.34	0.40	-1.62	2.52	-0.16	0.13	0.36	0.36
5	0.81	4.20	1.25	0.25	0.99	0.32	0.92	0.35	0.33	0.36	-1.62	2.42	-0.16	0.12	0.36	0.35
6	2.16	12.09	1.60	0.60	1.39	0.73	1.33	0.77	0.75	0.69	-0.62	3.61	0.07	0.21	0.74	0.53
7	3.19	27.76	1.80	0.97	1.60	1.13	1.54	1.19	0.93	0.90	-0.29	4.04	0.16	0.26	0.88	0.62
8	1.15	7.02	1.35	0.32	1.12	0.40	1.05	0.43	0.48	0.41	-1.14	2.77	-0.07	0.14	0.54	0.40
9	5.58	84.89	2.15	1.95	1.98	2.18	1.92	2.27	1.19	1.30	0.11	4.77	0.28	0.33	1.06	0.80
10	4.50	75.99	1.95	1.70	1.76	1.92	1.70	2.00	1.00	1.22	-0.29	4.86	0.18	0.32	0.90	0.80
r	0.9587		0.9713		0.9715		0.9716		0.9862		0.9809		0.9882		0.9825	
p	0.0001		0.0001		0.0001		0.0001		0.0001		0.0001		0.0001		0.0001	

X = Original variables T1 = $\sqrt{X}+1$; T2 = $\sqrt{X}+0.5$; T3 = $\sqrt{X}+0.375$; T4 = log (X + 1); T5 = log (X + (k/2)); T6 = log [log (X + 2)]; T7 = $X^{1-(b/2)}$; Var.: variance.

Métodos estadísticos no paramétricos y transformaciones de datos en estudios de poblaciones de plagas agrícolas. Al analizar datos provenientes de poblaciones de plagas agrícolas, regularmente se detecta que no cumplen los requerimientos teóricos para la aplicación del ANDEVA clásico. El uso de transformaciones Box-Cox y de métodos estadísticos no paramétricos resulta la alternativa más utilizada para resolver este inconveniente. En el presente trabajo se exponen los resultados de la aplicación de estas técnicas a datos provenientes de *Thrips palmi* Karny muestreadas en plantaciones de papa (*Solanum tuberosum* L.) en el período de incidencia de la plaga. Se utilizó la dócima χ^2 para la bondad de ajuste a la distribución binomial negativa y de independencia para investigar la relación entre los estratos de las plantas y los estados del insecto, se aplicaron siete transformaciones a los datos para satisfacer el cumplimiento de los supuestos básicos del ANDEVA, con las cuales no se logró eliminar la relación entre la media y la varianza, por lo que se aplicó el ANDEVA no paramétrico de Kruskal-Wallis para comparar las densidades poblacionales del insecto. Los resultados permitieron seleccionar el estado larval del insecto y el estrato medio de la planta como elementos claves para diseñar planes de muestreo para esta plaga.

Palabras clave: dócima Kruskal-Wallis, distribución binomial negativa, transformaciones Box-Cox, *Thrips palmi*, *Solanum tuberosum*.

LITERATURE CITED

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society-Series B 26:211-252.

Cabrera, A. 2002. Criterios estadísticos en la descripción del patrón espacial y diseño de muestreos para *Thrips palmi* Karny en papa. 99 p. Tesis doctoral. Centro Nacional de Sanidad Agropecuaria, Universidad Agraria de La Habana, La Habana, Cuba.

Cabrera, A., M. Suris, W. Guerra, y D.E. Nicó. 2008. Muestreo secuencial para la toma de decisiones de control de *Thrips palmi* en papa en Cuba. Manejo Integrado de Plagas y Agroecología (Costa Rica) 79-80:13-22.

Conover, W.J. 1999. Practical nonparametric statistics. 3rd ed. John Wiley & Sons, New York, USA.

Costello, M.J., and K.M. Daane. 1997. Comparison of sampling methods used to estimate spider (Araneae) species abundance and composition in grape vineyards. Environmental Entomology 26:142-149.

De Calzadilla, J., W. Guerra, y V. Torres. 2002. El uso y abuso de transformaciones matemáticas. Aplicaciones en modelos de análisis de varianza. Revista Cubana de Ciencia Agrícola 36:103-106.

Di Rienzo, J.A., F. Casanoves, M.G. Balzarini, L. González, M. Tablada, y C.W. Robledo. 2008. InfoStat versión 2008. Grupo InfoStat, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Córdoba, Argentina.

Eisenhart, C. 1947. The assumptions underlying the analysis of variance. Biometrics 3:3-21.

Fernández, A., A. Cabrera, y J.I. Durán. 2005. Patrón espacial y distribución dentro de la planta de *Pseudacysta perseae* Heidemann (Hemiptera: Tingidae) en aguacateros asociados al cultivo del cafeto. Revista de Protección Vegetal 20:173-178.

Font, H., V. Torres, M. Herrera, and R. Rodríguez. 2007. Fulfillment of the normality and the homogeneity of the variance in frequencies of accumulated measurement of the egg production variable in White Leghorn hens. Cuban Journal of Agricultural Science 41:207-211.

Gómez, M., C. Danglot, y L. Vega. 2003. Sinopsis de pruebas estadísticas no paramétricas. Cuándo usarlas. Revista Mexicana de Pediatría 70:91-99.

Liscano, A., y O. Domínguez. 2005. Distribución de inmaduros de *Antichloris viridis* Druce, 1884 en la planta de plátano (Musa AAB, subgrupo plátano, cv. Hartón) en el sur del lago de Maracaibo, Venezuela. Revista de la Facultad de Agronomía 22:240-353.

Menchaca, M.A. 1973. Método corto para el análisis de transformaciones. Revista Cubana de Ciencia Agrícola 7:145-149.

Miranda, C., A. Vasconcellos, and A. Bandera. 2004. Termites in sugar cane in Northeast Brazil: Ecological aspects and pest status. Neotropica Entomology 33:237-241.

Peña, S. de R.D. 1994. Estadística. Modelos y métodos. 2. Modelos lineales y series temporales. 745 p. Alianza Editorial, S.A., Madrid, España.

Santos, B., J. Gilreath, R. Arbona, y A. Pimentel. 2005. La estadística no paramétrica para el análisis e interpretación de estudios de plagas: alternativas al análisis de varianza. Manejo Integrado de Plagas y Agroecología (Costa Rica) 75:83-89.

Siegel, S., y N.J. Castellan. 1995. Estadística no paramétrica aplicada a las ciencias de la conducta. 4a ed. 437 p. Editorial Trillas, México.

Sokal, R.R., and F.J. Rohlf. 1995. Biometry. 776 p. 3rd ed. Freeman, New York, USA.

Southwood, T.R.E. 1995. Ecological methods with particular reference of the study of insect populations. p. 7-69. Chapman and Hall, London, UK.

Steel, R.G., e I.H. Torrie. 1992. Bioestadística: principios y procedimientos. 740 p. McGraw-Hill Interamericana, México.

Suris, M., y L. Plana. 2001. Distribución en la planta y en el campo de *Thrips palmi* (Thysanoptera: Thripidae) en papa de la variedad Desirée. Revista de Protección Vegetal 16(2-3):80-83.

Tinoco, O. 2008. Una aplicación de la prueba chi cuadrado con SPSS. Revista de la Facultad de Ingeniería Industrial 11:73-77.

Tobing, M.C. 2007. The vertical distribution and monitoring population of *Thrips palmi* Karny (Thysanoptera: Thripidae) on *Solanum tuberosum* (potato plant). Jurnal Biosains 18(2):1-8.

Varga, L., P.J. Fedor, M. Suvák, J. Kiseľák, and E. Atakan. 2010. Larval and adult food preferences of the poinsettia thrips *Echinothrips americanus* Morgan, 1913 (Thysanoptera: Thripidae). Journal of Pest Science 83:319-327.

Verghese, A., P.L. Tandon, and G.S. Prasada Rao. 1988. Ecological studies relevant to the management of *Thrips palmi* Karny on mango in India. Tropical Pest Management 34:55-58.