COMPUTACIÓN Y MATEMÁTICA APLICADA
APLLIED COMPUTATION AND MATHEMATICS

ARTÍCULO ORIGINAL

Diferentes métodos estadísticos para el análisis de variables discretas. Una aplicación en las ciencias agrícolas y técnicas

Different Statistical Methods for the analysis of discrete variables. An application in the agricultural and technical sciences

Magaly Herrera Villafranca[1], Caridad W. Guerra Bustillos[2], Lucía Sarduy García[1], Yoleisy García Hernández[1] y Carlos Enrique Martínez[3]

RESUMEN. El objetivo del presente trabajo fue evaluar tres métodos estadísticos para el análisis de variables discretas. La información empleada proviene de un experimento desarrollado en la Empresa Genética Camilo Cienfuegos, de la provincia de Pinar del Río en el período 2007-2008, relacionada con la producción de CT-115. Se analizaron tres muestreos, como caso de estudio se seleccionó el muestreo dos que comprendió los meses junio-julio 2007, se empleó un diseño completamente aleatorizado con tres tratamientos y 10 repeticiones. Las variables analizadas fueron: No. tallos, No. rebrotes, No. hojas totales/tallos, No. hojas totales/rebrotes, No. hojas secas/tallos y No. hojas secas/rebrotes. Se tuvo en cuenta el Análisis de Varianza paramétrico, su homólogo no paramétrico la dócima Kruskal–Wallis y Modelo Lineal Generalizado. Se verificó el cumplimiento de los supuestos teóricos del Análisis de Varianza, para la normalidad de los errores se utilizaron las dócimas Shapiro Wilk, Kolmogorov Smirnov y Lilliefors, la dócima de Shapiro Wilk fue la más robusta para detectar la falta de normalidad, para la homogeneidad de varianza se emplearon las dócimas de Bartlett y Levene, ambas obtuvieron resultados similares. Las variables se transformaron según raíz cuadrada, la cual no mejoró el cumplimiento de distribución Normal para la variable No. hojas secas/rebrotes. Los valores de probabilidad obtenidos mantuvieron el mismo criterio de decisión con respecto a Ho para ambas dócimas, la no paramétrica Kruskal- Wallis comparado con su homóloga paramétrica la dócima F de Fisher. Los criterios de bondad de ajuste utilizados en el Modelo Lineal Generalizado permitieron conocer los efectos de mejor ajuste, Se considero que este modelo es más flexible que el Análisis de Varianza paramétrico, pues las variables en estudio no requiere del cumplimientos de los supuestos teóricos básicos.

Palabras clave: Transformaciones de datos, ANAVA paramétrico y no paramétricos, Modelo Lineal Generalizado

ABSTRACT. The objective of this article was to evaluate three different statistical methods to conduct analyses of discrete variables. The information came from an experiment developed at the Camilo Cienfuegos Genetics Enterprise in the Pinar del Río province in 2007-2008 related to the CT-115 forrage production. A complete randomized design was used with three treatments and ten repetitions. The variables analysed were: number of stems, number of sprouts, total number of leaves/stem, total number of leaves/sprout, number of dried leaves/stem and number of dried leaves/sprouts. The parametric variance analysis and its homologous non-parametric, Kruskal-Wallis test and the Generalized Lineal Model were taken into account. The theoretical assumptions of the variance analyses to the test error normality were verified. The Shapiro Wilk test, Kolmogorov Smirnov and the Lilliefors test were used, Shapiro Wilk test was the most robust to detect lack of normality. For the variance homogeneity, the Bartlett and Levene test were used both with similar results. The variables were transformed with the square root transformation which did not improve the normal distribution adjustment to the variable number of dried leaves/sprout. The probability values maintained the same outcomes respect to Ho tests for the non-parametric test, compared with its homologous parametric F from Fisher test. The criteria of goodness of fit in the Generalized Linear Model permitted evaluating the best adjustment effects. It was considered that this model is more flexible than the parametric variance analyses because the variables under study did not require the theoretical assumptions fulfilment.

Keywords: Transformation of data, parametric ANOVA and non-parametric, Generalized Linear Model

Recibido 20/10/10, aprobado 10/12/11, trabajo 01/12, artículo original.

[1] M.Sc., Inv., Instituto de Ciencia Animal, Carretera Central km. 47½, San José de las Lajas, Mayabeque, Cuba, E-✉: mvillafranca@ica.co.cu

[2] Dr. C. Prof. Titular, Centro Universitario Municipal Güines, Mayabeque, Cuba

[3] Lic., Inv. Delegación Provincial de La Agricultura, Pinar del Río, Cuba.

INTRODUCCIÓN

Con frecuencia en la investigación científica en diversas ramas de la ciencia y en particular las agropecuarias, se presentan en los resultados variables de tipo discretas, asociadas con conteo y mucho de los casos son analizadas por métodos estadísticos en los que se hacen necesario que las variables cumplan determinados requisitos.

El Análisis de Varianza paramétrico requiere de supuestos de normalidad de los errores, homogeneidad de varianza, independencia de los errores y aditividad de efecto. Su comprobación se hace necesaria para sustentar la validez de análisis. La verificación de los supuestos subyacentes se realiza en la práctica a través de los predictores de los términos de error aleatorio que son los residuos aleatorios asociados a cada observación (Balzarini *et al.* 2008).

El incumplimiento de alguno de estos supuestos lleva a conclusiones erróneas como rechazar la hipótesis nula siendo verdadera o viceversa, lo que trae consigo resultados falsos en los experimentos que al materializarse en los sistemas de producción introducen pérdidas económicas y error en el proceso de toma de decisiones (Pérez *et al*. 2002).

Dentro del análisis de datos el Análisis de Varianza ha sido muy utilizado y dentro de este campo de la docimasia de hipótesis, las que con más frecuencia se utilizan son las pruebas paramétricas, ya que dados los supuestos que requiere su aplicación, resultan ser las más potentes (Cristo y Guerra, 2001). Autores como De Calzadilla *et al.* (2002) plantean que está no es la más efectiva cuando se incumplen algunos de los supuestos antes mencionado y sugieren el empleo de métodos no paramétricos como alternativas de análisis o el Modelo Lineal Generalizado. Según Fox (2005) este modelo se puede utilizar para variables continuas y discretas, el cual resulta de gran utilidad.

El objetivo del presente trabajo es evaluar tres métodos estadísticos para el análisis de variables discretas. El Análisis de Varianza paramétrico y su homólogo no paramétrico la dócima de Kruskal-Wallis y el Modelo Lineal Generalizado.

MÉTODOS

La información utilizada proviene de la Empresa genética Camilo Cienfuegos, de la provincia Pinar del Río, en el período 2007-2008, el estudio corresponde a tres muestreos realizados en áreas de pastoreos de la misma empresa, para el estudio se tomó como ejemplo el muestreo dos que comprende los meses junio- julio de 2007. Se empleó un diseño completamente aleatorizado con tres tratamientos y 10 repeticiones por cada uno. Las variables analizadas fueron: No. tallos, No. rebrotes, No. Hojas totales/tallos, No. hojas totales/rebrotes, No. hojas secas/tallos y No. hojas secas/rebrotes.

Verificación del cumplimiento de los supuestos teóricos del Análisis de Varianza paramétrico

Para la evaluación del supuesto de normalidad se utilizaron las dócimas de Shapiro Wilk, Kolmogorov-Smirnov y Kolmogorov-Smirno corregida por Lilliefors, propuestas por Diz (2008) y la homogeneidad de varianza se evaluó teniendo en cuenta la dócima, de Levene y la de Bartlett modificada por (Royston), propuesta por (Font, 2007).

Se verificó el cumplimiento de los supuestos teóricos del Análisis de Varianza teniendo en cuenta, las variables originales y para las transformadas se empleó la transformación raíz cuadrada, por ser variables de conteos.

Análisis de Varianza paramétrico y no paramétrico

Se aplicó Análisis de Varianza paramétrico según Diseño Completamente Aleatorizado y su homólogo no paramétrico la dócima de Kruskal-Wallis, donde se analizaron las probabilidades de error de tipo I.

Modelo Lineal Generalizado

Para el estudio también se tuvo en cuenta el Modelo Lineal Generalizado propuesto por Nelder y Wedderburn (1972). Se analizó la distribución de las variables y se observó que según una distribución de Poisson, con función de enlace log, en cada caso se consideró al tratamiento como efecto fijo. El modelo es el siguiente:

$$Yij = \beta_0 + \beta\tau_i + eij$$

donde:
Yij: valor esperados asociado a las variables dependientes analizadas;
$\beta_{0:}$ intercepto del modelo;
β: vector de parámetro desconocido asociado al efecto de los tratamientos;
τ_i : efecto de los tratamientos. (i = 1,2, 3);
eij: efecto del error aleatorio asociado a la j-ésima observación (j = 1,2……30).

La función de enlace es:

$$\eta(\mu) = \log(\mu)$$

donde:
$\eta(\mu)$: función que relaciona a la media con el predictor lineal;
log (μ): función de enlace asociada a la distribución Poisson.

Para la bondad de ajuste del modelo se tuvieron en cuenta los criterios de devianza y Chi cuadrado de Pearson.

Para el análisis de los supuestos teóricos se empleó el software Statistica StarSoft (2003). Para el Análisis de Varianza paramétrico y no paramétrico el software estadístico Infostat (2008). En el caso del Modelo Lineal Generalizado se empleó el software SAS (2007) versión 9.1.3 procedimiento (GENMOD, *Generalized Linear Model*).

RESULTADOS Y DISCUSIÓN

En la Tabla 1 se muestran los resultados del cumplimiento de los supuestos teóricos del Análisis de Varianza paramétrico para las variables originales y las transformadas. Al evaluar la homogeneidad de varianza por las dócimas de Bartllet y Levene, mostraron valores de de probabilidad mayores que 0,05, lo que indica el cumplimiento de este supuesto. Según Montgomery (2002), la dócima de Bartlett es sensible a la violación del

supuesto de normalidad, y para Correa y Castillo (2000) la dócima de Levene ofrece una alternativa más robusta (es aquella prueba que hace una estimación exacta de la probabilidad del errores de tipo I y II) que el procedimiento de Bartlett, ya que es poco sensible a la desviación de la normalidad, sin embargo en este estudio ambas dócimas presentaron un comportamiento similar en cuanto a la homogeneidad de varianza.

Al analizar el supuesto de normalidad de los residuos para las variables originales se observó que dócima Shapiro-Wilk fue la más sensible para detectar el incumplimiento de este supuesto, pues las variables No. hojas totales/tallos, No. hojas secas/rebrotes mostraron valores de probabilidad menores que 0,05, no así las demás variables analizadas que obtuvieron valores de probabilidad de error de tipo I superiores. En este mismo análisis, pero para las variables transformadas, se evidenció que la dócima Shapiro-Wilk fue más robusta para detectar la falta de normalidad, que las dócimas Kolmogorov Smirnov y Lillierfors, aunque se utilizó la transformación raíz cuadrada, esta no mejoró el cumplimiento de dicho supuesto para la variable No hojas secas/tallos. Según Espejo *et al.* (2001), las dócima Kolmogorov Smirnov y Lillierfors, es más conveniente cuando la variable que se analiza es de tipo continua u ordinal y es más efectiva para muestras grandes, plantean además que el empleo de la dócima Shapiro-Wilk, es mejor cuando las muestras que se analizan son de tamaño n< 50 observaciones. Teniendo en cuenta que el tamaño de muestra empleado en este estudio fue de 30 observaciones la dócima que se sugiere para evaluar la normalidad es la Shapiro-Wilk.

TABLA 1. Probabilidad de error de tipo I en las dócimas de homogeneidad de varianza y normalidad

Variables	Dócima homogeneidad de varianza		Dócimas de normalidad de los errores		
	Bartllet	Levene	Shapiro Wilk	Kolmogorov Smirnov	Lilliefors
No. tallos	0,11	0,36	**0,05**	0,20	0,10
No. rebrotes	0,48	0,56	0,37	0,20	0,20
No. hojas totales/tallos	0,57	0,58	***0,01***	0,20	0,10
No. hojas totales/rebrotes	0,16	0,12	***0,05***	0,20	0,15
No. hojas secas/tallos	0,17	0,08	0,16	0,20	0,05
No. hojas secas/ rebrotes	0,15	0,43	***0,01***	0,20	0,10
Raíz Cuadrada No. tallos	0,32	0,53	0,80	0,20	0,20
Raíz Cuadrada No. rebrotes	0,77	0,77	0,69	0,20	0,20
Raíz Cuadrada No. hojas totales/ tallos	0,79	0,80	0,05	0,20	0,05
Raíz Cuadrada No. hojas totales /rebrotes	0,87	0,85	0,19	0,20	0,20
Raíz Cuadrada No. hojas secas/tallos	0,18	0,12	0,19	0,20	0,10
Raíz Cuadrada No. hojas secas/ rebrotes	0,67	0,71	***0,01***	0,10	0,10

Los resultados obtenidos para los rangos de probabilidad (P) de de las dócimas F de Fisher y su homólogo no paramétrico la dócima Kruskal Wallis para las variables originales, se observó que de los 12 casos analizados el 83,3 % de eficiencia (10 valores de P que suman la diagonal principal) corresponden con los rangos de P donde coinciden ambas dócimas. En un estudio realizado por Guerra *et al.* (2000) comprobaron que el 96,1% de las variables mantuvieron el mismo criterio de decisión con selección a Ho para ambas dócimas.

Al realizar este mismo análisis para las variables transformadas se observó que los rangos de probabilidad de error de tipo I no variaron para ambas dócimas y mantuvieron el mismo criterio de decisión con relación a Ho. De Calzadilla (1999), encontró resultados similares. La comparación de lo valores de probabilidad se muestran en la Tabla 2.

TABLA 2. Comparación de los valores de probabilidad de las dócimas F de Fisher y su homólogo no paramétrico Kruskal-Wallis para las variables (originales o transformadas)

Dócima Kruskall-Wallis	Dócima de F de Fisher			Total
	< 0,01	0,01-0,05	> 0,05	
< 0,01	2	0	0	2
0,01-0,05	1	4	0	5
> 0,05	0	1	4	5
Total	3	5	4	12

Con el análisis realizado se obtuvo una aproximación de la eficiencia del ANOVA no paramétrica Kruskal- Wallis comparado con su homóloga paramétrica la dócima F de Fisher bajos las mismas condiciones.

En la Tabla 3 se presentan los resultados de Modelo Lineal Generalizado para las variables analizadas y el efecto del tratamiento, teniendo en cuenta los criterios de bondad de ajuste y la significación del efecto antes mencionado a partir de la distribución Poisson, con función de enlace log. A partir de los criterios analizados se observó que cuatro variables obtuvieron valores de desvianza y de Chi-cuadrado de Pearson cercanos a uno lo que indica que el modelo para esas variables presentó un buen ajuste. Al respecto Mora *et al.* (2007) reafirmaron que cuando los valores de desviación y Chi-cuadrado de Pearson dividido por los grados de libertad correspondientes muestran valores cercanos a 1, se evidencia un ajuste apropiado.

TABLA 3. Resultados de los criterios de bondad de ajuste de las variables en el Modelo Lineal Generalizado

Variables	GL	Valor de la desvianza	Valor X^2 De Pearson	Desvianza (Valor/GL)	$\chi 2$ de Pearson (Valor/GL)	Valor de P de tratamiento
No. tallos	27	25,95	27,53	0,96	1,01	0,08
No. rebrotes	27	13,05	13,08	0,48	0,48	0,05
No. hojas totales/ tallos	27	39,82	41,57	1,47	1,54	0,01
No. hojas totales/ rebrotes	27	19,71	19,92	0,73	0,74	0,001
No. hojas secas/ tallos	27	6,32	6,33	0,23	0,23	0,17
No. hojas seca/ rebrotes	27	45,29	37,87	1,68	1,40	0,11

En la Tabla 4 se observan los valores de los parámetros del modelo a partir del método de estimación de Máximo Verosimilitud obtenido del Modelo Lineal Generalizado, según Verde (2000) este método permite obtener valores estimados para los parámetros que maximizan la probabilidad de obtener el conjunto de datos en evaluación. Este método realiza la estimación haciendo máximo un valor de la función tal que cuando se evalúa la derivada a cero el resultado obtenido se corresponde con el de máximo verosimilitud. A su ve este se compara con el resto de los tratamientos y si difieren de cero, se plantea que existe un efecto significativo del tratamiento sobre las variables analizadas. Para este caso se observó que las variables No. tallos, No. hojas secas/tallos y No. hojas secas/rebrotes no presentaron diferencias significativas entre los tratamientos.

Este tipo de método estadístico es más flexible que el de Análisis de Varianza paramétrico, pues la variable respuesta no requiere del cumplimientos de los supuestos mencionados anteriormente, y el análisis se realiza teniendo en cuenta la función de enlace perteneciente a la distribución que adopten los datos que se estén analizando.

TABLA 4. Resultados del análisis de los parámetros estimados

Parámetros	GL	Estimadores	EE	X^2	Pr > X^2
No. tallos					
Intercepto	1	2,50	0,00	763,38	0,001
Tratamiento 1	1	-0,30	0,14	4,79	0,054
Tratamiento 2	1	-0,18	0,13	1,78	0,18
Tratamiento 3		0,00	0,00	0,00	
No. rebrotes					
Intercepto	1	2,14	0,11	389,29	0,001
Tratamiento 1	1	-0,49	0,17	7,79	0,005
Tratamiento 2	1	-0,21	0,16	1,66	1,66
Tratamiento 3		0,00	0,00	0,00	
No. hojas totales/ tallos					
Intercepto	1	3,79	0,05	6388,57	0,001
Tratamiento 1	1	-0,22	0,07	9,64	0,002
Tratamiento 2	1	-0,03	0,07	0,26	0,61
Tratamiento 3		0,00	0,00	0,00	
No. hojas totales/ rebrotes					
Intercepto	1	2,50	0,09	763,38	0,001
Tratamiento 1	1	0,63	0,11	32,54	0,001
Tratamiento 2	1	-0,09	0,13	0,43	0,51
Tratamiento 3		0,00	0,00	0,00	
No. hojas secas/ tallos					
Intercepto	1	2,40	0,09	643,07	0,001
Tratamiento 1	1	0,12	0,13	0,95	0,30
Tratamiento 2	1	-0,12	0,14	0,81	0,37
Tratamiento 3		0,00	0,00	0,00	
No. hojas secas/ rebrotes					
Intercepto	1	1,48	0,15	96,59	0,001
Tratamiento 1	1	-4,45	0,24	3,50	0,06
Tratamiento 2	1	-0,38	0,23	2,62	0,11
Tratamiento 3		0,00	0,00	0,00	

CONCLUSIONES

- Para este estudio se concluye que los resultados de las dócimas Bartlett y Levene, mantiene resultados similares en cuanto al supuesto de homogeneidad de varianza, pues en ambas dócimas obtienen valores de probabilidad por encima de 0,05. Para el análisis de normalidad de los errores la dócima Shapiro Wilk fue más conveniente de acuerdo al tamaño de muestra empleado n< 30 observaciones y al tipo de variables analizadas, por otra parte se observó que la transformación utilizada no mejoró el cumplimiento del supuesto de normalidad para la variable No. hojas secas/tallos. Los valores de probabilidad obtenidos en cuanto a los efectos de los tratamientos mantuvieron el mismo criterio de decisión con respecto a Ho para la dócimas no paramétrica Kruskal-Wallis comparado con su homóloga paramétrica F de Fisher. Los criterios de bondad de ajuste utilizados en el Modelo Lineal Generalizado permitieron conocer los efectos de mejor ajuste. El Modelo Lineal Generalizado es más flexible que el Análisis de Varianza paramétrico, pues las variable que se analizan no requieren del cumplimientos de los supuestos teóricos básicos, solamente se requiere conocer de la distribución que adoptan los datos analizados.

REFERENCIAS BIBLIOGRÁFICAS

BALZARINI, M.; A. DI RIENZO; F. CAZANOVES; L. GONZÁLEZ; M. TABLADA; W. GUZMÁN & W. ROBLEDO: *InfoStat software estadístico InfoStat versión 2008,* Manual de usuario, Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina, 2008.

CORREA, G. y A. CASTILLO: "Tamaño de muestra para aproximación de un estadístico a la distribución normal", *AGROCIENCIA*, 34(4): p.467–476, 2000.

CRISTO, M. y W. GUERRA: Comportamiento de las dócimas paramétricas respecto a las paramétricas en distribuciones no normales, *[en línea]* 2001*, Disponible en:* http://www.monografias.com/trabajos17/docimas-parametricas/docimas-parametricas.zip. *[Consulta: marzo 20 2011].*

DE CALZADILLA, J.: *Procedimientos de la Estadística No Paramétrica. Aplicaciones en las Ciencias Agropecuarias,* **Tesis (en opción al título de Máster en Matemática Aplicada a la Ciencias Agropecuarias),** La Habana, Cuba, 1999.

DE CALZADILLA, J., W. GUERRA y V. TORRES: "El uso y abuso de transformaciones matemáticas. Aplicaciones en modelos de análisis de varianza", *Revista Cubana de Ciencia Animal* 36(2): 103-106, 2002.

DIZ, R.: *Métodos para evaluar normalidad y homogeneidad de varianza. Relación con el tamaño de muestra,* 40pp., Trabajo de Investigación, Universidad de Granma, Bayamo, Cuba, 2008.

FONT, H.: *Estudio de precisión en la variable producción de huevos en gallinas White Leghorn*, 84pp., **Tesis (en opción al título de Máster en Producción Animal),** 2007.

FOX J.: *Generalized Linear Models: An introduction, York Summer Programme in Data Analysis*, Dpto. of Sociology McMaster, University in Hamilton, Ontario, Canada. 2005.

ESPEJO, I.; F. PALACÍN y A. LÓPEZ: *Inferencia estadística (Teoría y Problemas),* 267pp., Departamento de Estadística e Investigación Operativa, Universidad de Cádiz, España, 2001.

GRUPO INFOSTAT: *InfoStat software estadístico InfoStat versión 2008,* Manual de usuario, Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina, 2008.

GUERRA, C. W., J. DE CALZADILLA y V. TORRES: "Índice de eficiencia en relación con procedimientos de la estadística no paramétrica", *Revista Cubana de Ciencia Agrícola*. 34(1): 1-4, 2000.

MONGOTMERY, D. "Diseño y análisis de experimento", 2ª. Edn, Editorial Limusa, Medellín, Colombia, *Revista Ciencia e Investigación Agraria,* 34(2): 131-139, Chile, 2002.

MORA, F.; S. PERRET; C.A. SCAPIM; E. MARTINS e M. PAZ: *Variabilidad em el florecimiento de procedências Eucalyptus Cadocalyx en la Región de Coquimbo*, Brasil, 2007.

NELDER, J. A. & W. M. WEDDERBURN: "Generalized linear models", *Journal of the Royal Statistical Society*, 135(3): 370-384, 1972

PÉREZ, R; M. NODA; M. MORENO y E. PÉREZ: *Aplicación de la estadística en las diferentes etapas del ciclo de vida. Centro de Información y Gestión Tecnológica, Revista Trimestral, Año VIII, No. 2 Universidad de Holguín, Cuba [en línea] 2000, Disponible en:* http://www.ciencias.holguin.cu.2002/Junio/articulos/ARTl3.htm *[Consulta: agosto 7 2010].*

SAS: *User's guide statistics*, SAS Institute Inc., Cary, NC, USA, 2007.

STATSOFT, INC.: (STATISTICA (data analysis software system), version 6. www.statsoft.com. 2003.

VERDE, O.: "Comparación de métodos para análisis de datos binomiales en producción animal", *Revista Zootecnia Tropical,* 18(1):3-28, 2002.